Titre: Title:	Exploration et comparaison d'outils statistiques pour la prédiction du temps de guérison d'une plaie
Auteur: Author:	Violaine Mongeau-Pérusse
Date:	2017
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Mongeau-Pérusse, V. (2017). Exploration et comparaison d'outils statistiques pour la prédiction du temps de guérison d'une plaie [Mémoire de maîtrise, École Polytechnique de Montréal]. PolyPublie. https://publications.polymtl.ca/2765/

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/2765/
Directeurs de recherche: Advisors:	Nadia Lahrichi, & Louis-martin Rousseau
Programme: Program:	Maîtrise recherche en génie industriel

UNIVERSITÉ DE MONTRÉAL

EXPLORATION ET COMPARAISON D'OUTILS STATISTIQUES POUR LA PRÉDICTION DU TEMPS DE GUÉRISON D'UNE PLAIE

VIOLAINE MONGEAU-PÉRUSSE
DÉPARTEMENT DE MATHÉMATIQUES ET DE GÉNIE INDUSTRIEL
ÉCOLE POLYTECHNIQUE DE MONTRÉAL

MÉMOIRE PRÉSENTÉ EN VUE DE L'OBTENTION
DU DIPLÔME DE MAÎTRISE ÈS SCIENCES APPLIQUÉES
(GÉNIE INDUSTRIEL)
AOÛT 2017

UNIVERSITÉ DE MONTRÉAL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Ce mémoire intitulé :

EXPLORATION ET COMPARAISON D'OUTILS STATISTIQUES POUR LA PRÉDICTION DU TEMPS DE GUÉRISON D'UNE PLAIE

présenté par : MONGEAU-PÉRUSSE Violaine
en vue de l'obtention du diplôme de : Maîtrise ès sciences appliquées
a été dûment accepté par le jury d'examen constitué de :

M. AGARD Bruno, Doctorat, président
Mme LAHRICHI Nadia, Ph. D., membre et directrice de recherche
M. ROUSSEAU Louis-Martin, Ph. D., membre et codirecteur de recherche
M. DUHOUX Arnaud, Ph. D., membre

REMERCIEMENTS

J'aimerais tout d'abord remercier mes directeurs de recherche, Nadia Lahrichi et Louis Martin Rousseau, merci pour votre aide et votre soutien tout au long de ce projet. Un immense merci à Jonathan Vallée, VP Enginnering à Alaya Care. Merci à toute l'équipe de Alaya Care pour votre aide et vos encouragements, un merci plus spécifique à Pierre, Pau et David. En plus de votre aide, vous avez créé un espace agréable à travailler. Un immense merci aussi au professeur François Bellavance pour ses conseils et son expertise en data mining. Merci spécial à mes collègues du bureau, merci pour votre présence, votre appui et votre bonne humeur. Finalement un immense merci à mon amoureux, à ma famille et mes amis, sans vous je n'y serais pas arrivée. Merci à vous tous.

RÉSUMÉ

Présentement, dans la province de Québec, près de 40% du budget total du gouvernement est alloué au secteur de la santé. Avec le phénomène du constant vieillissement de la population, on suppose que ce chiffre ne cessera de s'accroître dans les prochaines années. En effet, nous constatons qu'actuellement près de 18% de la population est âgée de plus de 65 ans et les experts prévoient encore une augmentation de cette population pour toute la décennie à venir. Il est donc d'une importance cruciale de développer des outils ainsi que des méthodes efficaces pour diminuer les coûts en santé tout en conservant des soins optimaux pour les patients. Une option intéressante pour diminuer ces coûts est l'utilisation du service de soins à domicile. Les patients reçoivent des soins à domicile pour diverses raisons comme par exemple, la préparation des médicaments ou encore les soins des plaies.

Donc, une problématique importante touchant cette population spécifique des personnes âgées, concerne les plaies. Nous avons travaillé en collaboration avec une agence d'infirmières de soins à domicile provenant de l'Ontario. Sur l'ensemble de leurs patients, près de 40% reçoivent des soins en lien avec une problématique de plaies.

L'objectif de ce projet est de prédire le temps de guérison d'une plaie. Pour ce faire, nous avons utilisé des algorithmes d'intelligence machine. Ainsi, en ayant la possibilité de prédire correctement la durée de soins à donner pour un nouveau patient, le gestionnaire en charge pourra utiliser de façon plus judicieuse et efficiente les ressources infirmières de son équipe.

Pour réaliser ce projet, nous avons utilisé les données qu'a récoltées notre partenaire, la compagnie *AlayaCare*. L'agence d'infirmières emploie le logiciel qu'a créé *AlayaCare* pour inscrire les informations des patients. Après avoir anonymisées les données, il a été possible de les utiliser pour notre recherche. Nous avons donc tenté, en analysant les observations présentes dans la base de données obtenue, de prédire correctement le temps de guérison d'une nouvelle plaie chez un patient. Deux types de modèles de prédiction sont utilisés pour ce faire soit de régression, pour trouver un nombre de jours exact pour la variable cible ou encore la classification pour déterminer dans quel intervalle se trouve notre variable cible. Les méthodes appliquées sont la régression linéaire, la régression logistique, les arbres de décision et les forêts aléatoires.

De nombreuses modifications et analyses des bases de données pour ce projet furent réalisées pour rendre les données utilisables avec les algorithmes choisis. Par exemple, les valeurs manquantes ainsi qu'extrêmes et aberrantes ont été imputées ou encore supprimées. Nous avons aussi fait des analyses des variables disponibles.

Concernant nos résultats pour les modèles de classification, le taux de bonne classification se situe entre plus de 52% et plus de 86%. L'importante différence ici, existe pour plusieurs raisons. Entre autres, parce que nous avons ajouté des classes dépendant des essais effectués. Ainsi, il est plus difficile pour un modèle de prédire correctement lorsqu'il y a un plus grand nombre de classes. Pour les résultats des modèles de régression, le nombre de jours d'erreurs moyens se situe entre 20 et 33 jours. À ce stade, il est impossible d'affirmer ou d'infirmer si nos résultats sont satisfaisants. Or, nous avons été en mesure de répondre à notre objectif qui était de prédire le temps de guérison d'une plaie.

De nombreux essais ont été réalisés pour déterminer quel modèle présentait les meilleurs résultats. À chaque essai, nous avons modifié les paramètres de nos modèles pour trouver les meilleurs résultats. On remarque cependant que la base de données utilisée n'est pas optimale et que beaucoup d'erreurs semblent s'y être glissées, c'est une des principales limites de nos résultats. À plusieurs reprises, les données sont manquantes ou même erronées à certains endroits. Par exemple, dans plus de 10% des cas, il y a une valeur manquante pour la variable longueur de la plaie. De plus, de nombreux facteurs qui influencent la guérison d'une plaie ne sont pas présents dans la base de données telle que la présence de tabagisme ou encore d'alcoolisme. Une base de données plus complète c'est-à-dire avec un nombre moindre de valeurs manquantes et avec l'ensemble de variables explicatives nécessaires pour prédire la cible aurait pu nous donner de meilleurs résultats. Cependant, certaines lacunes peuvent aussi s'être glissées dans notre méthodologie. Par exemple, nous aurions pu modifier davantage de paramètres dans nos modèles ainsi il aurait peut-être été possible d'obtenir des résultats plus concluants. Malgré tout, force est de constater que nos résultats peuvent quand même apporter une certaine amélioration dans le fonctionnement de la compagnie d'infirmières à domicile. En effet, en obtenant des informations plus complètes au sujet des patients, le gestionnaire pourra améliorer sa gestion pour les horaires des infirmières.

ABSTRACT

Currently, in the province of Quebec, nearly 40% of the government's total budget is allocated to the health sector. With the phenomenon of the constant aging of the population, it is assumed that it will continue to increase in the coming years. Indeed, we note that currently about 18% of the population is over 65 years of age and experts are still predicting an increase in this population for the whole of the decade to come. It is therefore crucial to develop tools and effective methods to reduce health costs while maintaining optimal patient care. An attractive option to reduce these costs is the use of the home care service. Patients receive home care for a variety of reasons, such as medication preparation and wound care.

Therefore, an important problem affecting this specific population of the elderly concerns wounds. We worked with an agency of home care nurses from Ontario. Almost 40% of all patients receive wound care.

The objective of this project is to predict the healing time of a wound. We used machine leaning algorithms. Thus, with the ability to accurately predict the length of care for a new patient, the manager in charge will be able to make better use of his team's nursing resources in a more judicious and efficient manner.

To realize this project, we used the data collected by our partner *AlayaCare*. The nursing agency uses the software *AlayaCare* created to record patient information. After anonymizing the data, it was possible to use the data for our research. We have therefore attempted, by analyzing the observations present in the database obtained, to correctly predict the healing time of a new wound in a patient. Two types of prediction models are used to do this either regression, to find an exact number of days for the target variable or classification to determine in which interval our target variable is. The methods applied are linear regression, logistic regression, decision trees and random forests.

Numerous modifications and analyzes of the databases for this project were carried out to make the data usable with the chosen algorithms. For example, the missing values as well as the extreme and outliers were imputed or eliminated. We also analyzed the available variables.

Concerning our results for the classification models, the rate of good classification is between more than 52% and more than 86%. The important difference here exists for several reasons. Among other things, it is because we have added classes depending on the tests performed. Thus, it is more difficult for a model to predict correctly when there are more classes. For

the results of the regression models, the average number of days of errors is between 20 and 33 days. At this point, it is impossible to say or deny whether our results are satisfactory. However, we were able to meet our goal of predicting the healing time of a wound.

Numerous trials have been conducted to determine which model has the best results. At each test, we modified the parameters of our models to find the best results. We note, however, that the database used is not optimal and that many errors seem to have slipped into it, it is one of the main limitations of our results. On several occasions, the data is missing or even erroneous in some places. For example, in more than 10% of cases, there is a missing value for the wound length variable. In addition, many factors that influence the healing of a wound are not present in the database such as the presence of smoking or alcoholism. A more complete database with fewer missing values and the set of explanatory variables necessary to predict the target could have given us better results. However, some gaps may also have crept into our methodology. For example, we could have modified more parameters in our models so it might have been possible to obtain more conclusive results. Nevertheless, it is clear that our results can still bring some improvement in the operation of the home nursing company. Indeed, by obtaining more complete information about the patients, the manager will be able to improve its management for the schedules of the nurses.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ANNEXES

CHAPITRE 1 INTRODUCTION

Les dépenses du gouvernement du Québec en santé ne cessent d'augmenter depuis plusieurs années. Au Québec, pour l'année 2017-2018, le gouvernement provincial prévoit dépenser 38% de son budget seulement pour le secteur de la santé, ce qui équivaut à 40,36 milliards de dollars pour l'année. On note ici une augmentation de 4% par rapport à l'année précédente [37]. Selon l'institut canadien d'information sur la santé, une hausse des dépenses au niveau de la santé a eu lieu de 1998 à 2008. Cette hausse est attribuable à divers facteurs tels que les changements démographiques (croissance et vieillissement) et à la technologie. Durant cette période, les dépenses totales en santé ont augmenté annuellement d'en moyenne 7,4% [39]. En effet, depuis 2001, la population de 65 ans et plus est passée de 13% de la population générale à 18,1%. Il y a donc 542 275 personnes de plus dans cette catégorie d'âge depuis 15 ans [11]. Concernant les coûts reliés à cette catégorie d'âge, on remarque que pour l'année 2013, les dépenses moyennes par personne âgée de 65 ans et moins étaient d'environ 2 000$, tandis que ce chiffre augmentait de façon significative pour une moyenne de 10 000 $ pour les personnes âgées entre 65 et 85 ans [11].

Il est donc d'une importance capitale de trouver des moyens, et ce, le plus rapidement possible afin de diminuer et de contrôler les coûts reliés au système de la santé. Un changement prometteur est l'augmentation des soins à domicile. De tels soins s'avèrent plus économiques que les soins hospitaliers [28]. De plus, les risques d'infection sont beaucoup plus faibles en soins à domicile qu'en centre hospitalier et finalement les patients eux-mêmes préfèrent demeurer à la maison plutôt qu'à l'hôpital [9]. En effet, une réduction du nombre d'hospitalisations, ainsi qu'une probabilité moindre de placement dans un centre, sont les résultats d'une augmentation des services de soins à domicile [9]. Pour toutes ces raisons, il est primordial d'encourager les soins et services à domicile. Les raisons qui peuvent mener une personne à avoir recours aux soins à domicile sont nombreuses telles que le besoin de préparation des médicaments ou encore le changement de pansement [2].

Dans le cadre de ce projet, nous avons collaboré avec un partenaire industriel *AlayaCare*, spécialisé dans la conception de logiciels intégrés pour les soins à domicile. Parmi les problématiques soulevées, l'optimisation de la gestion des soins de plaies occupe une place importante, autant par le nombre de plaies que par les coûts engendrés. La clientèle ciblée par ces soins est âgée en moyenne de 66 ans. Considérant le vieillissement de la population et les données concernant l'incidence des plaies chez les personnes âgées, l'optimisation de la gestion des soins de plaie est donc un sujet actuellement important. En effet, au Québec, plus de 7% des

usagers de ces services auraient déclaré présenter une plaie [20]. Dans le cadre de ce projet, la compagnie *AlayaCare* nous a fourni les données nécessaires représentant les observations de la base de données d'une agence d'infirmières en Ontario. Selon ces données, 39,45% de leurs clients recevant des soins, présentent une ou encore plusieurs plaies. Il s'agit donc bien d'une problématique répandue dans cette agence d'infirmières.

Notre objectif est de prédire le temps de guérison d'une plaie à partir de la première visite d'une infirmière au domicile du patient. Ce projet fait partie d'un projet de plus grande envergure qui a comme objectif de diminuer les coûts associés aux plaies chez les patients nécessitant des soins à domicile. De nombreux projets peuvent être réalisés pour diminuer ces coûts, cependant pour ce mémoire nous nous concentrons sur le projet de prédire le temps de guérison d'une plaie à la première visite du patient. Nous utilisons seulement les données de la première visite, car lorsqu'un patient se présente avec une nouvelle plaie, il est impossible de savoir comment va évoluer la plaie. Ainsi, si nous utilisions les données de visites subséquentes, notre modèle serait biaisé.

Pour répondre à notre objectif qui est de prédire le temps de guérison d'une nouvelle plaie, nous allons utiliser plusieurs algorithmes. Nous serons donc en mesure de déterminer en combien de jours la plaie du patient devrait guérir. Ainsi, en sachant d'avance la durée approximative de soins qu'un patient nécessitera, nous pourrons planifier de façon optimale la demande de ressources. Nous serons aussi en mesure de mieux gérer l'effectif infirmier nécessaire et ainsi des coûts pourront être épargnés. En effet, il sera possible de savoir si une infirmière supplémentaire est nécessaire pour couvrir la demande des prochaines semaines ou bien si la plaie devrait guérir en moins de quelques jours, donc que l'agence est capable de répondre aux besoins avec son effectif actuel. Il ne devrait donc pas avoir de surplus ou encore de manque d'infirmière à combler, ce qui pourrait être économique pour l'agence.

Afin de prédire le temps de guérison d'une plaie, deux options s'offrent à nous. Plus spécifiquement, la première consiste à prédire le nombre exact de jours. Puisque nous tentons de prédire une valeur continue, les algorithmes qui sont les mieux adaptés sont des algorithmes de régression. Une seconde option consiste à prédire des intervalles de guérison. Bien que ce soit un cas particulier du premier et moins précis, le résultat peut être satisfaisant pour un décideur dans la mesure où obtenir un nombre de jours peut être difficile. Cette option consiste à prédire un intervalle de temps. Nous désirons que le modèle nous prédise si oui ou non, la valeur fait partie de cet intervalle. Ainsi, dans ce cas, les algorithmes utilisés sont de type classification. Plus généralement, les algorithmes employés dans ce projet sont la régression linéaire et logistique, les arbres de décision ainsi que les forêts aléatoires. Pour les arbres et les forêts, il est possible de modifier le modèle pour le rendre de type régression ou

encore de classification.

Peu importe l'algorithme utilisé, un ensemble de variables pouvant expliquer la durée de guérison d'une plaie est utilisé. Le choix de cet ensemble est inspiré d'une part de la littérature, et d'autre part des informations contenues dans la base de données. Parmi ces facteurs, nous retrouvons l'âge ou encore la présence de pathologies autres.

Grâce au logiciel qu'a créé *AlayaCare*, lorsqu'une infirmière de cette agence se présente au domicile d'un patient, elle remplit, sur sa tablette électronique, les informations pertinentes. Les données se retrouvent par la suite dans trois fichiers distincts : soit un premier fichier contenant toutes les informations relatives aux plaies, un deuxième avec les informations générales du patient et le dernier fichier contenant les plans de soins des patients. Il est important de regrouper ces fichiers pour être en mesure d'utiliser les modèles. Il s'agit d'un défi important puisqu'à certains endroits les données sont non structurées. Il a donc été nécessaire d'utiliser du *text mining* pour décortiquer toutes les informations et aller chercher ce qui est pertinent à notre projet. Finalement, notons que toutes les données ont été anonymisées pour respecter la confidentialité du patient et qu'un certificat éthique a été obtenu.

Comme dans tout projet de *data mining*, nous allons utiliser une approche qui se décompose en plusieurs étapes. Deux différentes approches sont présentées dans le chapitre 2.

Le mémoire a été divisé en plusieurs parties, en s'inspirant fortement des phases de tout projet de ce genre. Le prochain chapitre fera état d'une revue de littérature concernant divers aspects importants pour ce projet. Par la suite, il sera question de la méthodologie générale. Cette section sera séparée en plusieurs parties et toutes les étapes pertinentes à un projet de ce genre seront énoncées. Le cas d'étude suivra, et dans cette section, nous reprenons toutes les étapes décrites dans la méthodologie générale et nous les appliquons à notre projet. Les modèles prédictifs utilisés seront aussi expliqués dans ce chapitre. Suite au cas d'étude, les résultats des modèles seront énoncés et une discussion ainsi qu'une analyse approfondie seront présentées. Le mémoire se termine avec la conclusion.

CHAPITRE 2 REVUE DE LITTÉRATURE

Cette revue de littérature nous guidera dans le cheminement du projet. Elle est séparée selon trois grands thèmes, soit l'organisation des soins de plaies, les algorithmes utilisés pour réaliser notre projet et les projets semblables déjà effectués dans la littérature.

2.1 Organisation des soins de plaies

Les plaies représentent une problématique importante chez les personnes âgées [54]. En effet, tel qu'énoncé dans l'introduction, près de 40% des patients recevant des soins par l'agence d'infirmières avec laquelle nous travaillons présentent une ou plusieurs plaies. Actuellement, dans cette agence d'infirmières, il n'y a pas de protocole à suivre pour déterminer le nombre de visites ainsi que le moment d'effectuer une visite. L'horaire des visites est déterminé selon le jugement de l'infirmière en charge du patient. Par exemple, si le patient a une plaie qui nécessite un changement de pansement à chaque jour, l'infirmière devra y aller quotidiennement, et ce jusqu'au moment où elle détermine qu'elle peut espacer ses visites. De plus, dans certains cas, le médecin prescrit exactement les fréquences des soins qu'il désire, donc l'infirmière doit suivre les indications de celui-ci.

Selon une infirmière expérimentée en soins de plaies, actuellement il est possible qu'une infirmière prédise si deux patients ayant des caractéristiques très différentes auront une guérison semblable ou non. En effet, pour ce faire, elle s'inspirerait des facteurs influençant le temps de guérison d'une plaie. Toujours selon cette infimière, les facteurs influençant la guérison d'une plaie s'avèrent donc un aspect important à connaître.

De nombreux articles ont donc été étudiés pour obtenir une liste, la plus exhaustive possible de tous les facteurs influençant la guérison d'une plaie. Puisque la majorité des auteurs séparent en trois catégories les facteurs, c'est ce qui a été fait dans la revue de littérature présentée. Les trois catégories étant les facteurs systémiques, les facteurs locaux ainsi que les facteurs organisationnels. Les facteurs systémiques agissent sur l'état de santé du patient et non directement sur la plaie. Tandis que les facteurs locaux, eux, concernent des caractéristiques directement liées à la plaie [22]. Les facteurs organisationnels ne sont pas des caractéristiques du patient ni de sa plaie, mais bien de l'environnement qui l'entoure [21].

Le tableau 2.1 représente un résumé des facteurs qui influencent la guérison d'une plaie tel qu'énoncé dans plusieurs études [5, 23, 25, 38, 47, 49, 58].

Chaque facteur énuméré dans le tableau 2.1 influence la guérison d'une plaie et ce à divers

Tableau 2.1 Résumé des facteurs influençant la guérison d'une plaie

Facteurs systémiques	Facteurs locaux	Facteurs organisationnels
Alimentation et hydratation déficientes	Infection	Infirmière spécialisée
Déficit en oxygène	Plaie chronique	Équipe multidisciplinaire
Stress	Superficie plaie	Continuité des soins
Mauvaise perception sensorielle	Présence corps étranger	Évaluation complète
Âge, sexe	Hématome	Utilisation des données probantes
Obésité	Site de la plaie	
Diabète, HTA	Tissus nécrotiques	
Tabagisme, consommation alcool	Pression sur la plaie	
Maladie auto-immune	Hydratation de la plaie	
Croyances culturelles	Vascularisation de la plaie	
Taux d'hormones sexuelles	Type de plaie	

niveaux. Par exemple, pour le facteur systémique *sexe* de la personne, plusieurs études démontrent qu'une plaie nécessite plus de temps à guérir lorsqu'il s'agit d'une femme. Une des causes qui explique cette variation est l'œstrogène, hormone présente en beaucoup plus grande quantité chez la femme [59]. De plus, concernant un facteur local, s'il y a présence d'une infection au niveau de la plaie, l'inflammation sera prolongée, ce qui retardera la guérison du site. En effet, la présence de biofilms, bactéries complexes, empêche la guérison [25]. Aussi, la présence d'une infirmière spécialisée accélère la guérison d'une plaie. Effectivement, celle-ci est davantage en mesure de gérer efficacement l'évolution d'une plaie et de déceler les complications plus rapidement pour ainsi diminuer le temps de guérison de la plaie [47].

2.2 Algorithmes utilisés dans ce projet

Pour ce projet, nous avons décidé d'utiliser du *data mining* pour atteindre notre objectif. Depuis de nombreuses années, le *data mining* est utilisé par plusieurs compagnies de marketing à travers le monde [6]. Le *data mining* combine des analyses statistiques, de l'intelligence artificielle ainsi que des technologies pour extraire des relations et des patrons dans des immenses bases de données [56]. Pour réaliser un projet de *data mining*, divers outils peuvent être utilisés tels que la segmentation où le but est de découvrir des groupes ainsi que des structures dans les bases de données étudiées ou encore des outils d'association pour identifier des règles, par exemple, à des fins de *marketing* [6]. Les outils qui nous intéressent dans ce projet concernent un autre type d'outils soit les méthodes de régression ainsi que de classification. Dans ces cas, on s'intéresse à la prédiction.

2.2.1 Méthodes utilisées dans les projets de *data mining*

Avant de débuter un tel projet, il faut déterminer quelle méthode nous devons privilégier pour avoir une approche rigoureuse et complète. Diverses méthodes sont utilisées dans les projets de *data mining*. En effet, plusieurs auteurs ont énoncé des méthodologies à suivre. Dans cette revue de littérature, deux méthodes sont énoncées.

Premièrement, il y a les phases d'un projet de *data mining* selon le modèle de *Cross industry standard process for data mining*. Selon eux, un projet de ce genre est un processus itératif et adaptatif [10]. Il a été développé en 1996 par un groupe d'experts [6]. La figure 2.1 illustre cette méthode.

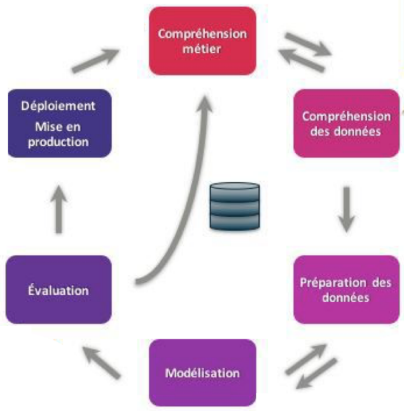

Figure 2.1 Méthode CRISP illustrée. ® Business & Decision

Chacune des étapes présentées à la figure 2.1 possède des objectifs précis et l'on peut retourner à une étape précédente au besoin. Il ne s'agit donc pas d'un processus linéaire. Dans cette traduction, l'auteur parle de compréhension du métier, mais dans d'autres articles il nomme cette étape compréhension de la problématique. Il faut, entre autres dans cette étape, déterminer nos objectifs. Par la suite, il y a les étapes de compréhension et préparation des données où l'on travaille sur les données présentes. La préparation des données est l'un des aspects les plus importants dans un projet de *data mining* [10]. Ensuite, il s'agit de la modélisation, c'est à cette étape que l'analyste choisit les modèles et qu'il les testera à plusieurs reprises en modifiant les paramètres pour déterminer quel est le meilleur modèle. Toujours à cette étape, il doit déterminer les critères d'évaluation pour comparer les modèles entre eux. Arrive ensuite l'étape de l'évaluation. À ce stade, il faut s'assurer que nos résultats sont conformes avec les critères de réussite commerciale. Finalement, la dernière étape selon cette méthode est le déploiement, c'est-à-dire utiliser les nouvelles connaissances pour apporter des améliorations dans les processus [10].

Deuxièmement, Berry et Linoff ont eux aussi élaboré une méthode qui se nomme le cercle vertueux du *data mining* [7]. Comme son nom l'indique, il ne s'agit toujours pas d'un processus linéaire, mais dans ce cas d'un cercle.

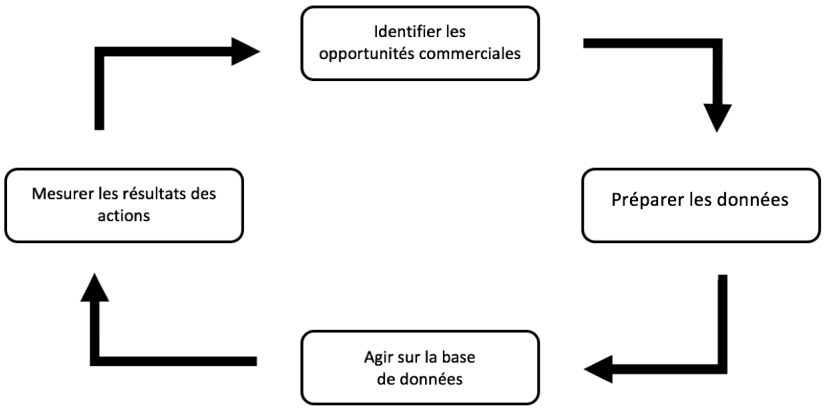

Figure 2.2 Cercle vertueux du *data mining* selon Berry et Linoff [7]

On remarque que dans le cercle vertueux du *data mining*, illustré à la figure 2.2, il y a moins d'étapes que dans le modèle présenté à la figure 2.1. La première étape, soit l'identification des opportunités commerciales consiste à identifier le domaine d'étude. De plus, il faut déterminer l'objectif du projet à ce stade. Par la suite, l'analyste doit faire une analyse des données présentes et les regrouper. À la troisième étape, l'analyste met en œuvre les techniques de *data mining* choisi. Finalement, tout comme l'étape d'évaluation selon la méthode CRISP, il faut mesurer les résultats des actions [45].

Les prochains paragraphes font un survol des outils de *data mining* que nous allons utiliser dans ce projet.

2.2.2 Régression linéaire multiple

Ce modèle est l'outil statistique le plus fréquemment utilisé pour l'étude de données ayant plusieurs variables indépendantes [8]. La régression linéaire est utilisée pour prédire un problème que l'on suppose linéaire. On dit qu'il s'agit d'une régression linéaire multiple lorsqu'il y a plusieurs variables explicatives. La variable cible est de nature continue [24]. Ce type de modèle tente de tracer une droite sur un certain nombre de points, et ce en ayant la plus petite somme des distances de chaque point à la droite [24].

La figure 2.3 montre un exemple d'une régression linéaire simple. Il s'agit de la même chose

pour une régression linéaire multiple, seulement il y a plusieurs dimensions.

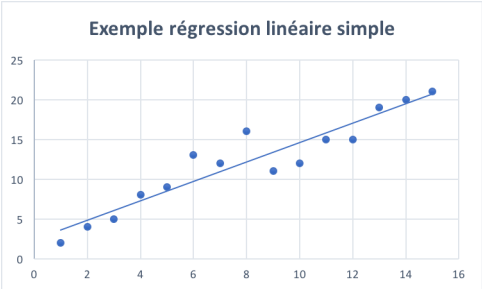

Figure 2.3 Exemple d'une régression linéaire simple

L'équation 1 est l'équation générale de la régression linéaire multiple.

$$(1) y_i = (b_0 + b_1 x_1 + b_2 x_2 + ... + b_n x_n) + e_i \quad \forall i = 1...n$$

où x_i est une variable explicative, b_i est son coefficient bêta, b_0 est l'ordonnée à l'origine, y_i est la variable cible et e_i est l'erreur que réalise le modèle pour chaque valeur de y.

2.2.3 Régression logistique multinomiale

La régression logistique prédit non pas une valeur continue, mais plutôt si la valeur fait partie ou non d'une catégorie. Dans la mesure où nous avons plusieurs catégories, il faut utiliser la régression logistique multinomiale [44]. L'équation de la régression logistique ressemble beaucoup à celle de la régression linéaire, mais en y ajoutant une transformation logarithmique [44].

$$(2) P(Y) = \frac{1}{1 + e^{-}(b_0 + b_1 x_1 + b_2 x_2 + ... + b_n x_n)}$$

où $P(Y)$ est la probabilité que Y arrive, e est la base des logarithmes naturels et les autres valeurs sont les mêmes qu'à l'équation de la régression linéaire.

2.2.4 Arbre de décision

Les arbres de décision peuvent être autant de l'ordre de la régression, donc la variable cible doit être continue ou encore de l'ordre de la classification, ainsi la variable cible est catégorielle. Dans les deux cas, les modèles ont pour but d'estimer la valeur de la variable cible. Un avantage des arbres est que le modèle final est très visuel. Nous pouvons donc facilement

démontrer aux personnes intéressées pourquoi notre modèle présente une telle prédiction, alors que c'est impossible à faire avec d'autres algorithmes tel que les forêts aléatoires. Le principe des arbres est très simple, il s'agit de « diviser l'ensemble des données d'apprentissage successivement en sous-groupes, selon les valeurs prises par les variables explicatives qui, à chaque étape, discrimine le mieux la variable cible » [6]. Voici quelques avantages de cette méthode :

1. Les règles sont simples
2. Les règles sont facilement interprétables
3. Peu de traitement de données à faire

La figure 2.4 montre un exemple d'un arbre de décision, la variable cible dans cet énoncé est la présence ou l'absence d'une infection. La variable explicative est la rougeur. On remarque que lorsque la caractéristique rougeur est présente, 80% du temps, il y a une infection. Il s'agit d'un exemple très simple, mais le même principe s'applique pour des problèmes plus complexes.

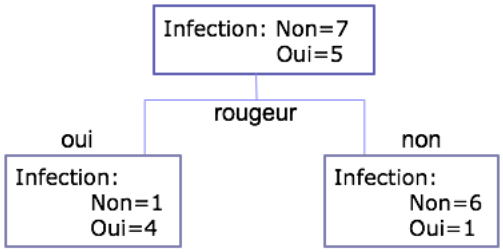

Figure 2.4 Exemple d'un arbre de décision simple

2.2.5 Forêt aléatoire

La forêt aléatoire est une méthode d'ensemble. En résumé, le modèle crée un nombre X d'arbres de décision, en utilisant à chaque fois un échantillon des données. Pour chaque ensemble de données, il y a une sélection aléatoire de variables explicatives. Ensuite, toutes les prédictions sont stockées pour chaque observation d'origine et finalement la prédiction finale que conclura la forêt aléatoire sera la prédiction ayant été obtenue le plus souvent par les arbres [36]. Les méthodes d'ensembles, tels que les forêts aléatoires, donnent souvent de meilleurs résultats et sont plus stables. Cependant, leur point négatif est qu'il n'est plus possible de les visualiser [6]. Pour les forêts aléatoires de type classification, la variable cible est divisée en classe. Pour les forêts de type régression, il s'agit d'une variable cible continue.

2.2.6 Critères d'évaluation

Dans tous les projets de *data mining*, il est primordial d'avoir des critères d'évaluation pour être en mesure de comparer nos modèles prédictifs.

Critères d'évaluation pour les modèles de classification

Pour la classification, les critères suivants sont régulièrement utilisés dans la littérature :

1. Sensibilité
2. Spécificité
3. Précision
4. Taux de bonne classification

La matrice de confusion sert à évaluer les modèles. Les valeurs qu'on y retrouve nous aident par la suite à calculer les critères d'évaluation soit la sensibilité, la spécificité, la précision et le taux de bonne classification.

Le tableau 2.2 représente la matrice de confusion générale.

Tableau 2.2 Matrice de confusion

	Prédiction : Dans cet intervalle	Prédiction : Non présent dans cet intervalle
Observé : Dans cet intervalle	Vrai positif (VP)	Faux négatif (FN)
Observé : Non présent dans cet intervalle	Faux positif (FP)	Vrai négatif (VN)

Pour bien comprendre la matrice de confusion, il est important d'avoir une définition claire des valeurs qui s'y retrouvent. Le tableau 2.3 donne donc ces définitions.

Tableau 2.3 Définition des valeurs de la matrice de confusion

	Définition
Vrai positif	Observation prédite dans le bon intervalle
Faux positif	Observation prédite dans l'intervalle alors qu'elle ne devrait pas y être
Vrai négatif	Observation prédite correctement dans un autre intervalle
Faux négatif	Observation qui auraient dû être dans cet intervalle

Dans certains projets, il y a plusieurs catégories, et non seulement une variable cible binaire, il est donc un peu plus compliqué de calculer ces critères. En effet, le seul critère qui concerne l'ensemble des catégories est le taux de bonne classification. Pour la sensibilité, la spécificité et la précision, il faut les calculer pour chaque intervalle. On se retrouve avec plusieurs valeurs pour la sensibilité, la spécificité et la précision. Il faudrait donc réaliser la matrice de confusion pour chaque catégorie distincte. Dans le but de comparer les résultats des modèles entre eux, une moyenne pour chaque critère doit être calculée.

La sensibilité représente la proportion de vrais positifs qui ont été correctement identifiés par le modèle [1].

$$(3) Sensibilité = \frac{VP}{VP + FN}$$

La spécificité, quant à elle, est la proportion de vrais négatifs qui ont aussi été correctement identifiés par le modèle [1].

$$(4) Spécificité = \frac{VN}{VN + FP}$$

Concernant la précision, il s'agit de la proportion de vrais positifs sur l'ensemble des valeurs positives.

$$(5) Précision = \frac{VP}{VP + FP}$$

Le taux de bonne classification est la proportion des bonnes prédictions sur l'ensemble des valeurs utilisées.

$$(6) Taux\,de\,bonne\,classification = \frac{VP + VN}{n}$$

où n est le nombre total de valeurs.

Critères d'évaluation pour les modèles de régression

Les modèles de régression utilisent des critères d'évaluation différents de ceux de classification. De plus, ils diffèrent régulièrement d'un article à un autre. L'erreur quadratique moyenne est cependant un critère d'évaluation connu pour les modèles de régression. L'erreur quadratique moyenne compare la valeur prédite à la valeur observée [4]. Voici l'équation de cette erreur :

$$(7) Erreur\, quadratique\, moyenne = \sqrt{\frac{1}{n}\sum (x_i - y_i)^2} \quad \forall i = 1...n$$

où n est le nombre total d'observations, x_i la valeur actuelle de l'observation i et y_i la valeur prédite de l'observation i.

Lorsque nous sommes en mesure de comparer les résultats selon nos différents critères d'évaluation, il faut aussi s'assurer qu'il n'y a pas de sur-apprentissage dans nos modèles.

2.2.7 Techniques pour vérifier la présence de sur-apprentissage

Le sur-apprentissage se définit comme le fait d'apprendre parfaitement les observations présentes. Le modèle connaît donc de mémoire la variable cible à prédire pour toutes les observations. Ainsi lorsqu'on présente une nouvelle observation au modèle, celui-ci n'est plus en mesure de prédire correctement la valeur [6]. Il est donc primordial d'utiliser des moyens pour vérifier la présence de sur-apprentissage dans nos modèles. Plusieurs techniques s'offrent à nous pour vérifier une telle situation.

Un cas typique de sur-apprentissage survient avec les modèles d'arbre de décision où le paramètre de la profondeur peut être très grand. Dans ce cas, le modèle créera des branches qui divisent toutes les données précisément et donc pour les données, il y aura un taux de mauvaise classification très bas, mais lorsqu'un nouveau cas sera présenté le modèle ne prédira pas correctement [17].

Deux techniques sont utilisées dans la vaste majorité des articles trouvés pour vérifier la présence de sur-apprentissage. Il s'agit de l'utilisation de deux fichiers distincts et la validation croisée.

Concernant la première technique, le principe est simple. Il suffit de séparer l'ensemble des données en deux fichiers, soit le fichier d'apprentissage et le fichier test. Ainsi toutes les données ne sont pas toujours présentes. Le fichier d'apprentissage sert à construire le modèle et celui test à valider le modèle. Un différent pourcentage est utilisé dépendant des auteurs pour le nombre d'observations dans chaque fichier. Habituellement, entre 60-70% des observations sont dans le fichier d'apprentissage et de 30-40% dans le fichier test. On doit par la suite déterminer un pourcentage jugé acceptable entre les deux fichiers, ce pourcentage dépend de plusieurs facteurs. Un des facteurs importants pour déterminer ce seuil est la taille de l'échantillon. S'il y a une différence supérieure à la limite fixée, cela signifierait qu'il y a du sur-apprentissage dans le modèle.

L'autre technique disponible est la validation croisée. Elle est souvent utilisée lorsqu'il y a

un nombre restreint d'observations disponibles [6]. Le mécanisme est simple, le fichier est divisé en n groupes. Par la suite, on entraînera le modèle en utilisant n - 1 groupes et on testera, c'est-à-dire on fera la prédiction sur l'échantillon restant. Le modèle sera entraîné n fois. Finalement, le taux de bonne classification ou encore le taux d'erreurs est la moyenne de tous les essais [6]. La figure 2.5 illustre la validation croisée avec cinq sous-ensembles. On remarque qu'à l'itération 1, le modèle est entraîné avec les quatre premiers sous-ensembles et à chaque itération l'échantillon change pour le test. En gris, il s'agit donc des sous-ensembles pour apprendre au modèle et en blanc, c'est le sous-ensemble pour tester.

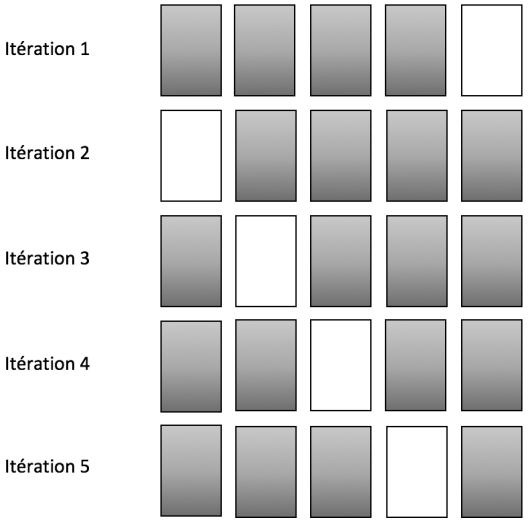

Figure 2.5 Validation croisée avec n = 5

2.3 Projets similaires dans la littérature

Une recherche en profondeur en lien avec le temps de guérison d'une plaie et le *data mining* a été effectuée. Or, nous n'avons pas été en mesure de trouver des articles précis concernant la prédiction des temps de guérison des plaies. Puisque notre projet se déroule dans un contexte de soins à domicile, une recherche plus précise à propos de la prédiction de problèmes de santé chez les patients demeurant à domicile a été réalisée, cependant, encore une fois nous n'avons pas été en mesure de trouver des articles dans cette veine. Nous avons donc fait une recherche sur des projets semblables à celui réalisé dans ce mémoire, mais en touchant à n'importe quelle pathologie de santé.

Les articles sélectionnés pour cette revue de littérature sont ceux qui, selon, nous semblaient les plus pertinents. Ils suivent tous une démarche générale de *data mining* en détaillant

toutes les étapes. Sans avoir cherché spécifiquement pour un type de maladie, on remarque que certaines pathologies reviennent plus régulièrement dans les articles de *data mining*. Le sujet le plus présent, selon nos recherches, en *data mining* médical est la prédiction de la présence ou l'absence d'une problématique de type cardiaque [14, 31, 43, 55]. La deuxième problématique la plus présente selon nos recherches est la prédiction de tout type de cancer [16, 34, 46].

Un autre aspect important en *data mining* touche la taille de la base de données. Deux articles contenaient une base de données avec plus de 150 000 observations [16, 46]. Pour tous les autres articles, cette valeur varie entre 250 observations et 7000 observations [14, 15, 26, 31, 34, 40, 41, 43, 52, 55]. Les articles ne spécifient pas si ces valeurs sont pour la base de données initiale ou finale, c'est-à-dire s'ils ont fait leurs modèles avec ce nombre d'observations ou s'ils ont dû supprimer des observations pour diverses raisons. Ainsi, nous ne connaissons pas le nombre final d'observations utilisées avec les modèles.

Nous remarquons que dans la majorité des cas, la variable cible est catégorielle et non de régression. En effet, ces études cherchent à prédire l'absence ou la présence d'un diagnostic X ou encore la présence dans un intervalle [15, 31, 41, 46, 52].

Pour les variables explicatives, c'est-à-dire les facteurs influençant la variable cible, plusieurs articles repérés utilisent le jugement d'experts et/ou la littérature pour faire le choix des variables à exploiter. Dans certains articles, ils prennent d'abord toutes les variables explicatives disponibles dans la base de données, ensuite un expert médical retire les variables non pertinentes [34, 55]. Cependant dans d'autres cas, ils utilisent directement les variables dictées par les experts et la littérature [15, 31, 41, 46, 52].

De plus, parfois les auteurs ne donnent pas d'explications sur la technique utilisée pour le choix des variables explicatives [16, 40, 43]. Seulement Palaniappan [43] suggère dans ses recommandations qu'il aurait pu augmenter le nombre de variables explicatives s'il avait demandé l'avis d'experts dans le domaine. Finalement, l'article de Heikes [26] possède 18 variables explicatives, or dans l'article, il est écrit que plusieurs variables explicatives ont été rejetées, car elles n'apportaient pas de gain informationnel significatif.

Concernant les modèles utilisés, la majorité des articles proposent plus d'un modèle et ils effectuent une comparaison entre eux. Le seul modèle présent dans tous les articles est l'arbre de décision. L'autre modèle le plus utilisé est les réseaux de neurones [14, 15, 16, 34, 40, 41, 43, 46, 55]. La régression logistique est utilisée pour quatre articles retenus pour notre revue de littérature [16, 26, 40, 41]. Les autres modèles retrouvés dans certaines études sont les séparateurs à vaste marge, les forêts aléatoires et les modèles de classification Naïve bayésienne.

Trois techniques ont été utilisées par les articles sélectionnés pour s'assurer qu'il n'y ait pas de sur-apprentissage. Les trois techniques sont les suivantes : validation croisée, fichiers distincts et aucune technique. En effet, certains auteurs ont seulement indiqué dans leur conclusion que leur modèle présentait un fort risque de sur-apprentissage, car aucune technique n'a été employée pour éviter et vérifier la présence de sur-apprentissage [14, 31, 52]. Les articles [15, 16, 26, 40, 41, 46, 55] ont quant à eux employé la validation croisée avec soit 5 ou 10 sous-ensembles. Finalement les articles [34, 43] utilisent la technique des deux fichiers distincts.

Pour tous les articles trouvés, utilisant des algorithmes de classification, les mêmes critères d'évaluation ont été pris en considération [15, 16, 26, 40, 41, 46, 55]. Le critère principal était le taux de bonne classification, c'est-à-dire le pourcentage du temps que le modèle a prédit correctement une observation. Les critères secondaires identifiés sont entre autres, la sensibilité, la spécificité et la précision.

Pour ce qui est des résultats, pour le taux de bonne classification, cela varie entre 75% et 100% avec une moyenne de 78,9%. Dans la majorité des cas étudiés pour cette revue de littérature, il s'agit de l'arbre de décision qui a donné les meilleurs résultats [16, 26, 34, 41, 43, 46, 52, 55]. Il faut se rappeler qu'ici les auteurs prédisent une variable binaire et non plusieurs classes ou encore une valeur continue.

Tel qu'énoncé au début de cette section, nous n'avons pas été en mesure de trouver d'articles directement en lien avec la prédiction en *data mining* et le temps de guérison d'une plaie. Cependant, un article fut découvert en lien avec des tentations de prédire, mais en n'utilisant pas d'algorithme de prédiction.

En effet, l'article [48] a fait une étude clinique en lien avec les ulcères diabétiques du pied. 276 patients ont participé à cette étude. Les auteurs cherchaient à déterminer quel pourcentage de diminution de la superficie de la plaie à quatre semaines était déterminant d'une guérison complète à 12 semaines. Ils ont donc réussi à prédire si en 12 semaines une plaie serait guérie en utilisant le facteur de la guérison à 4 semaines.

CHAPITRE 3 MÉTHODOLOGIE GÉNÉRALE

Tout projet de *data mining* suit une approche précise tel qu'énoncé dans la section 2. Pour ce projet, nous avons décidé de suivre la méthode du modèle de *Cross Industry Standard Process for Data Mining* [10]. Dans cette section, une description claire et transférable du projet est produite. Quatre grandes sections sont expliquées dans le chapitre. Il s'agit de la compréhension des données, de la préparation des données, de la modélisation et finalement de l'évaluation. Il faut prendre note que toutes les informations disponibles dans cette section proviennent de la méthode choisie [6, 10].

3.1 Compréhension des données

À cette étape, un recensement des données présentes dans nos bases de données est fait. Il est important de comprendre comment toutes les données sont récoltées, et ce dans le but de pouvoir les utiliser convenablement. De plus, une recherche approfondie est réalisée pour s'assurer que toutes les données nécessaires sont disponibles. La variable cible est la variable que l'on souhaite prédire et les variables explicatives sont les caractéristiques indépendantes qui nous aideront à prédire la cible. Par exemple, la variable cible pourrait se définir comme suit : présence ou absence d'une crise cardiaque chez une personne et une des variables explicatives pourrait être la présence d'engourdissement dans son bras gauche. Dans cette étape, il y aura une phase de description des données où la quantité de données ainsi que la qualité des données seront analysées.

Une analyse descriptive des variables présentes est importante pour connaître les variables ainsi que pour être en mesure de bien les préparer. De plus, une brève description de chaque variable est nécessaire. Il faut comprendre ce que la variable signifie et apprendre tout ce qu'on peut sur celle-ci. Dans cette étape, il y aura aussi une analyse des valeurs manquantes ainsi qu'aberrantes et extrêmes.

Une valeur manquante implique qu'il existe une valeur pour cette observation, mais qu'elle n'a pas été prise ou bien qu'il n'y a aucune valeur correspondante [6].

Une valeur aberrante se définit comme « une valeur erronée, causée par une erreur de saisie, une erreur de calcul, une mauvaise mesure ou une fausse déclaration » [6].

Les valeurs extrêmes sont des valeurs éloignées des autres valeurs dans la population [6]. De plus, une valeur extrême n'est pas nécessairement aberrante. Par exemple une valeur rare telle qu'être âgé de 100 ans pourrait être un extrême, mais sans être une valeur aberrante.

Certaines équations peuvent être utilisées pour détecter les extrêmes. On peut utiliser les équations suivantes :

Équation pour limite inférieure :

$$(8) inf = \mu - n\sigma$$

Équation pour limite supérieure :

$$(9) sup = \mu + n\sigma$$

où μ est la moyenne, σ l'écart-type et n le nombre d'écart-type utilisé.

3.1.1 Analyse et réduction du nombre de variables explicatives

Dans le but d'utiliser le plus efficacement possible les modèles, il est important de faire une sélection de variables explicatives. Cette sélection se divise en deux étapes distinctes. Premièrement, il faut déterminer quelles sont les variables explicatives qui ont une corrélation avec la variable cible. Deuxièmement, il faut sélectionner les variables explicatives qui sont fortement corrélées entre elles et en garder une seule [35].

Avant de débuter, voici l'équation pour calculer le coefficient de corrélation [13].

$$(10) Coeff.corrélation = \frac{\sum(x-m)\cdot(y-n)}{\sqrt{\sum(x-m)^2}\cdot\sqrt{\sum(y-n)^2}}$$

où x et y sont les valeurs de la variable cible ainsi que de la variable explicative choisie respectivement. De plus, m et n sont les moyennes de ces variables.

Variables explicatives corrélées à la variable cible

Plusieurs méthodes sont possibles pour déterminer si une variable cible est corrélée à la variable cible. Voici une liste des méthodes possibles. Une brève description de ces méthodes sera par la suite donnée.

1. Retirer les variables explicatives une à la fois et vérifier si les résultats de prédiction sont meilleurs
2. Vérifier les coefficients de corrélation
3. Effectuer un nuage de points entre chaque variable explicative et la variable cible
4. Calculer les coefficients de détermination

Pour la méthode 1, chaque variable explicative est retirée une après l'autre pour vérifier si le taux d'erreur diminue en retirant cette variable. En utilisant le coefficient de corrélation, la variable la moins fortement corrélée avec la variable cible est retirée en premier et ainsi de suite pour se rendre jusqu'à la plus corrélée. Si le taux d'erreur diminue en retirant la variable, alors cette variable est retirée définitivement, si le contraire se produit, la variable est rajoutée au modèle.

Pour la méthode 2, il faut vérifier le coefficient de corrélation. Un coefficient de corrélation près de 0 signifie que la variable explicative n'influence pas la variable cible. S'il est négatif, c'est qu'il influence négativement la variable cible et inversement lorsqu'il est positif. De plus, pour avoir une forte corrélation entre deux variables, le coefficient devrait se situer entre 0,5 et 1 ou entre -0,5 et -1 [18].

Pour la méthode 3, il suffit de faire un nuage de points entre chaque variable explicative et la variable cible. On pourrait être en mesure de constater une certaine tendance.

Finalement pour la méthode 4, il s'agit de calculer le coefficient de détermination soit le r^2. Le coefficient de détermination représente le coefficient de corrélation, mais au carré. De meilleurs résultats devraient être obtenus de cette manière. Si nous avions un r^2 de 0,10, cela signifierait que seulement 10% de la variance de la variable choisie est justifiée par la corrélation. La corrélation dans ce cas est donc très mauvaise [42].

Après l'analyse de toutes les valeurs et des variables, il faut préparer les données et faire les modifications nécessaires.

3.2 Préparation des données

La préparation des données est l'étape la plus volumineuse de la méthodologie [10]. C'est à cette étape que la base de données utilisée par les modèles sera constituée. De plus, de nombreuses modifications des données ainsi que des variables doivent être apportées aux les données.

3.2.1 Constitution de la base de données

Lorsque nous connaissons l'emplacement de toutes les données pertinentes à notre étude, il est ensuite le temps de constituer la base de données. Parfois les données peuvent être à divers endroits, cependant pour qu'un modèle de prédiction fonctionne les données doivent toutes se retrouver dans le même fichier. Il est donc important de constituer la base de données qui sera utilisée par la suite.

3.2.2 Rejet de variables explicatives

Certaines variables explicatives doivent être complètement rejetées, et ce pour diverses raisons. La liste suivante énumère les raisons pour lesquelles un analyste pourrait rejeter certaines de ces variables.

1. La variable a toujours la même valeur (constante), elle n'apporte donc aucune valeur prédictive
2. La variable présente un grand nombre de valeurs manquantes
3. La variable ne nous apporte pas de nouvelles informations
4. La variable est fortement corrélée avec une autre variable (colinéarité)
5. La variable est redondante

3.2.3 Modification des valeurs manquantes

Une autre modification importante à faire concerne les valeurs manquantes. Pour plusieurs algorithmes, il est important de s'assurer qu'il n'y ait pas de valeur manquante sinon le modèle ne fonctionnera pas. Plusieurs méthodes s'offrent à nous pour ne plus avoir de valeur manquante. Voici une liste de ces méthodes.

1. Imputer par la valeur 0
2. Imputer par la valeur moyenne
3. Imputer par distribution
4. Supprimer les observations avec données manquantes

Parfois, l'analyste est en mesure de savoir facilement par quelle méthode remplacer une valeur manquante. Par exemple, dans une base de données contenant les prix de maisons, il pourrait y avoir la caractéristique de présence (1) ou absence (0) d'une piscine. Dans ce cas, on peut supposer que s'il y a une valeur manquante pour cette variable, c'est parce que la caractéristique n'est pas présente. L'analyste pourrait donc remplacer toutes les valeurs manquantes de cette variable explicative par 0. Pour les autres variables explicatives, l'analyste testera toutes les options et utilisera finalement celle qui lui donne les meilleurs résultats de prédiction. De plus, tel que présenté dans la section 3.2.2, l'analyste peut rejeter complètement une variable si celle-ci présente trop d'observations manquantes. Encore une fois, il doit tester son modèle de prédiction et vérifier avec quelle méthode il obtient les meilleurs résultats.

3.2.4 Modification des valeurs aberrantes

Plusieurs options s'offrent à nous lorsqu'il y a présence de valeurs aberrantes. L'analyste choisira donc la meilleure méthode qui convient à ses données. Voici une liste des méthodes possibles qu'on peut utiliser lorsqu'on détecte la présence de valeurs aberrantes.

1. Corriger par la vraie valeur si possible
2. Supprimer l'observation
3. Rejeter les variables ayant un grand nombre de valeurs aberrantes
4. Imputer par la moyenne
5. Accepter d'avoir quelques valeurs aberrantes

3.2.5 Modification des valeurs extrêmes

Dans certains algorithmes, tel que la régression linéaire, les extrêmes peuvent fortement influencer nos résultats, il faut donc les modifier. Pour effectuer ces modifications, nous pouvons utiliser les mêmes options qu'à la section 3.2.4.

De plus, en utilisant les équations décrites à la section 3.1, nous sommes en mesure de déterminer les limites inférieures et supérieures pour chaque variable continue. Il est ainsi possible de trouver l'intervalle où les données doivent se situer. Plusieurs essais sont effectués avec différentes valeurs de n et la valeur qui donne les meilleurs résultats de prédiction est gardée.

En ayant les limites inférieures et supérieures pour chaque variable continue, il est possible de trouver l'intervalle où les données doivent se situer. Toutes les valeurs ne faisant pas partie de cet intervalle sont donc considérées comme des extrêmes.

3.2.6 Création de nouvelles variables

Il est souvent nécessaire de créer de nouvelles variables. Parfois certaines variables ne sont pas utiles telles quelles, mais après quelques modifications, elles peuvent le devenir. Par exemple, si l'on soustrait la date d'aujourd'hui à la date de naissance, on peut trouver l'âge du client. Cette variable peut être plus utile pour un projet.

3.2.7 Création de classes pour les modèles de classification

Dans le même ordre d'idée, il faut parfois créer des classes pour les modèles de classification. Ces classes sont déterminées par l'analyste. C'est à lui de choisir quelles classes lui semblent

les plus pertinentes et lesquelles donnent les meilleurs résultats de prédiction. L'analyste peut décider de créer autant de classes qu'il le souhaite. Il doit cependant s'assurer qu'il y ait assez d'observations dans chaque classe. On peut utiliser la matrice de confusion pour bien visualiser où se retrouvent les données. Le tableau 3.1 est un exemple d'une matrice de confusion simple.

Tableau 3.1 Exemple de matrice de confusion

Prédit/ Observé	1 an à 25 ans	26 ans et plus
1 an à 25 ans	2000	50
26 ans et plus	100	200

3.2.8 Normalisation des variables continues

Une autre modification importante à apporter est de normaliser les données. On évite ainsi « qu'une variable soit plus importante qu'une autre dans l'algorithme à cause de son unité de mesure, sa moyenne et sa variance »[6]. Pour normaliser, plusieurs méthodes sont possibles. Voici une équation utilisée dans plusieurs projets :

$$(11) x_i = \frac{v_i - \mu}{\sigma} \quad \forall i = 1...n$$

où μ est la moyenne, σ l'écart-type, v_i la valeur initiale et n le nombre d'observations.

La normalisation doit être réalisée sur les données des variables explicatives continues et non pour les variables binaires ou avec des classes. Ces variables n'ont donc plus d'unité de mesure. De plus, il est essentiel de noter la moyenne et l'écart type qui sont utilisés pour chaque variable. En effet, si de nouvelles observations s'ajoutent, il faudra normaliser les valeurs de ces observations avec les mêmes constantes (moyenne et écart type).

3.3 Modélisation

C'est dans cette étape que les modèles seront choisis. De nombreux modèles sont disponibles dans la littérature. Il demeure à l'analyste de déterminer quels modèles sont les plus pertinents pour le projet. De plus, il faut déterminer une procédure pour évaluer et tester notre modèle. Une liste des paramètres à modifier et une marche à suivre pour tester nos modèles devraient être réalisées dans cette section [10]. L'analyste doit aussi décider quels critères d'évaluation sont les plus pertinents à utiliser.

3.4 Évaluation des résultats

À chaque début de projet, l'analyste doit déterminer des objectifs commerciaux ainsi que des critères de réussite commerciale. À l'étape d'évaluation des résultats, il doit s'assurer que ses objectifs ont été atteints, et ce en répondant aux critères de réussite qu'il s'était fixés.

CHAPITRE 4 CAS D'ÉTUDE

Pour le cas d'étude, nous nous sommes inspirés de la méthode de CRISP. Les différentes étapes et le schéma représentant cette méthode sont illustrés à la section 2. Dans cette partie du travail, nous focalisons sur la compréhension ainsi que sur la préparation des données et sur les modèles à utiliser. Il y aura la détection des valeurs manquantes, extrêmes, aberrantes et des modifications seront réalisées pour rendre la base de données fonctionnelle afin de l'utiliser dans les modèles de prédiction. De plus, nous nous concentrons sur les étapes 2, 3 et 4 de cette méthode soit la compréhension des données, la préparation des données ainsi que la modélisation.

4.1 Compréhension des données

Tel qu'énoncé dans le chapitre 3, l'étape de compréhension des données englobe plusieurs aspects importants. Les prochaines sous-sections feront donc état de tous ces points tels que l'analyse des données et des variables.

4.1.1 Caractéristiques des bases de données utilisées

Dans le projet actuel, les données se trouvent sur différentes bases de données. Effectivement, la base de données principale contient seulement les variables relatives à la plaie traitée. Une autre base de données contient uniquement certaines données démographiques telles que l'âge et le sexe de la personne et finalement dans le fichier des plans de soins, plusieurs autres informations supplémentaires s'y retrouvent.

Lorsqu'une infirmière se présente au domicile du patient, elle complète les champs demandés dans le logiciel d'*AlayaCare*. Dans le logiciel, les infirmières remplissent des champs prédéfinis. Par exemple, il est écrit « cocher » si le patient est atteint de diabète. Il est donc facile de savoir quelle est l'information entrée, c'est ainsi pour les deux premières bases de données. Cependant, pour les plans de soins, il s'agit d'une case où elles peuvent écrire tout ce qui leur semble nécessaire. Les informations se retrouvent par la suite dans trois fichiers *.csv* différents. Le tableau 4.1 fait état d'un résumé des informations incluses dans chaque base de données.

La base de données principale est l'endroit où la majorité des données ainsi que des variables se retrouvent. Dans cette base de données, il y a 177 991 lignes, et chaque ligne signifie qu'une visite a eu lieu pour une plaie spécifique. Pour une même journée, un patient peut donc avoir

Tableau 4.1 Caractéristiques des bases de données utilisées

Base de données	Informations contenues dans la base de données
1. Principale	Toutes les informations relatives à la plaie
2. Générale	Âge, sexe et diagnostics
3. Plan de soins	Information non structurée, mais jugée pertinente par l'infirmière

plusieurs visites à son dossier, il faut comprendre qu'il s'agit de la même visite, mais les soins visaient une plaie différente. Lors de chaque visite, l'infirmière remplit des champs dans le logiciel qu'*AlayaCare* a créé. Sur l'ensemble des visites, il y a 8381 plaies différentes, ce que nous estimons en moyenne à 21 visites par plaie pour l'ensemble du traitement.

4.1.2 Définition des variables disponibles

Il est très important de bien comprendre toutes les variables disponibles pour obtenir de bons résultats dans un tel projet. En effet, une analyse rigoureuse est nécessaire pour choisir comment préparer chacune des variables. Voici une brève description de toutes les variables disponibles. Il s'agit des variables que nous croyons pertinentes à la prédiction du temps de guérison d'une plaie et qui sont disponibles dans nos bases de données.

Sexe

Tel que mentionné dans la revue de littérature, le sexe de la personne est un facteur qui influence le temps de guérison d'une plaie. Il s'agit d'une variable binaire soit la valeur 1 pour une femme et 0 pour l'homme. De plus, près de 52% de tous les patients recevant des soins de plaie sont des hommes.

Type de plaie

Cette variable nous indique de quel type est la plaie. La plaie ne guérira pas de la même façon en raison de sa physiopathologie. En effet, certains types de plaies risquent davantage de s'infecter que d'autres [60]. Le type de plaie n'est pas une variable, car nous avons décidé de séparer nos résultats par type de plaie et donc chaque modèle a été effectué sur des fichiers contenant seulement les plaies du même type. Cependant pour certains tests, nous avons gardé l'ensemble de plaies et donc dans ces cas-là, la variable type est une variable qui peut nous servir dans la prédiction. Il n'a pas été possible de tester la prédiction sur tous les sous-ensemble de plaies, car il n'y avait pas assez de données disponibles. Les modèles

prédictifs ont été faits sur les plaies les plus fréquentes dans la base de données. Nous avons choisi de sélectionner seulement les types de plaies ayant un minimum de 900 observations. Le tableau 4.2 énumère les quatre plaies les plus communes ainsi que leur fréquence et leur prévalence dans la base de données.

Tableau 4.2 Les quatre types de plaies les plus fréquentes

Type de plaie	**Fréquence**	**% sur l'ensemble des plaies**
Plaie opératoire (1)	2398 plaies	28,6%
Trauma (2)	1061 plaies	12,6%
Autres (3)	1059 plaies	12,6%
Plaie de pression (4)	957 plaies	11,4%

Les fréquences diminueront significativement, en raison des valeurs manquantes et des extrêmes entre autres.

Statut de la plaie

Une plaie peut être nouvelle ou récurrente. Le statut de la plaie sera donc de 1 si la plaie est nouvelle et de 0 si elle est récurrente. Près de 28% des plaies étaient récurrentes et donc la majorité des plaies étaient nouvelles.

Acuité

L'acuité peut prendre cinq valeurs différentes. La plaie peut être aiguë, chronique, à guérison lente, de maintenance ou encore inconnue. Puisque nous ne voulions pas de mot dans notre base de données, nous avons remplacé ces valeurs par des numéros de 1 à 5 tel qu'indiqué dans le tableau 4.3. Ce même tableau indique aussi la fréquence de ces variables.

Tableau 4.3 Valeurs et fréquences pour l'acuité

Variable	**Numéro utilisé**	**Pourcentage sur l'ensemble de plaies**
Aiguë	1	68%
Chronique	2	16%
Guérison lente	3	8%
Maintenance	4	4%
Inconnu	5	4%

Emplacement

La variable emplacement signifie la localisation de la plaie, soit l'endroit où la plaie se retrouve sur le corps. Cette variable est très importante, car c'est un des facteurs influençant le plus la guérison. Par exemple, une plaie au talon prendra plus de temps à guérir qu'une plaie sur le bras, entre autres à cause de la pression [27]. Le tableau 4.4 donne les valeurs que nous avons utilisées pour chaque emplacement. L'emplacement le plus fréquent est *Genou-cheville* avec près de 15% de toutes les plaies.

Tableau 4.4 Valeurs utilisées pour chaque emplacement

Variable	**Numéro utilisé**	**Variable**	**Numéro utilisé**
Coccyx	1	Genou-cheville	8
Ischium	2	Hanche-trochanter	9
Fesses	3	Hanche-genou	10
Autre	4	Hanche-cheville	11
Pied	5	Dos	12
Talon	6	Bras-épaule	13
Milieu du dos	7	Orteil	14

Longueur, largeur, profondeur et superficie

Les variables longueur, largeur et profondeur signifient les dimensions initiales mesurées par l'infirmière. Finalement la superficie est la multiplication entre la largeur et la longueur de la plaie. Toutes les mesures sont en centimètres. Toutes ces variables sont essentielles pour la prédiction du temps de guérison d'une plaie. Plus une plaie est profonde ou encore large, plus la guérison sera lente [32]. Il s'agit ici de valeur continue. Le tableau 4.5 donne un aperçu des fréquences ainsi que des moyennes pour les dimensions de la plaie. Les moyennes, les minimums et maximums donnés sont pour la base de données complète avant d'avoir effectué toutes les modifications, il y a donc des valeurs aberrantes et extrêmes.

Tableau 4.5 Fréquence et moyenne des dimensions de la plaie

Variable	**Min**	**Max**	**Moyenne**
Superficie	0 cm	2850 cm	15.66 cm
Longueur	0 cm	120 cm	3.37 cm
Largeur	0 cm	76 cm	2.03 cm
Profondeur	0 cm	45 cm	0.67 cm

Tunnel

La variable tunnel est de type binaire. On dit qu'il y a présence d'un tunnel lorsqu'il y a une extension de la plaie initiale. Une valeur 1 signifie la présence d'un tunnel. Le tunnel peut s'étendre dans toutes les directions. Souvent ce tunnel est causé par une infection [3]. La présence d'une telle caractéristique risque d'augmenter le temps de guérison de la plaie. Seulement 3% de toutes les plaies recensées présentent cette caractéristique.

Espace sous-jacent

Il s'agit de déterminer s'il y a destruction tissulaire se situant sous la peau intacte près du pourtour de la plaie [33]. Tout comme pour le tunnel, il s'agit d'une variable binaire où il y a un 1 s'il y a présence de la caractéristique et 0 sinon. Près de 4% des plaies ont cette caractéristique.

Quantité d'exsudat

La variable ici nous indique la quantité d'exsudat. Les données possibles pour cette caractéristique sont : minime (0), modéré (1), abondant (2) et copieux (3). L'exsudat est un liquide qui s'écoule de la plaie. Cela peut parfois indiquer la présence d'une infection. 81% du temps, la valeur minime est indiquée. Les numéros dans les parenthèses représentent la valeur prise dans la base de données et c'est ainsi pour les prochaines variables aussi.

Odeur

L'évaluation de l'odeur est aussi un facteur important. Trois valeurs sont possibles : aucune (0), légère (1) et modérée (2). Différentes raisons peuvent expliquer la présence d'odeur, certaines plus bénignes que d'autres [29]. 89% du temps, aucune odeur n'est recensée pour les plaies.

Sensibilité de la peau

La perte de sensibilité près de la plaie est signe qu'il y a possiblement une lésion au niveau d'un nerf [57]. Heureusement, dans la majorité de nos cas, il n'y a pas de perte de sensibilité (0). Parfois, la sensibilité est diminuée (1) ou elle peut être complètement absente (2). 81% du temps, il n'y a aucune perte de sensibilité pour une plaie.

Douleur

Une gestion inefficace de la douleur peut mener à une guérison plus lente. La douleur au site d'une plaie peut être due à diverses causes telles que des lésions aux vaisseaux sanguins ou encore une infection [19]. Dans la base de données, la douleur est inscrite avec une échelle de 0 à 10, 0 étant aucune douleur et 10 une douleur très intense. Dans 64% des cas, aucune douleur n'est présente.

Âge

L'âge est un autre facteur qui joue grandement sur la guérison d'une plaie, et ce de la naissance à 65 ans. En effet, avec l'âge, dans la majorité des cas, la durée de guérison augmente. Ce retard serait causé par une réponse inflammatoire altérée [53]. Or, après 65 ans, le temps de guérison serait le même, peu importe l'âge. L'âge moyen, dans nos données, est de 66 ans et la médiane est de 68 ans.

Diabète

Pour de nombreuses raisons, la guérison d'une plaie chez une personne atteinte de diabète sera plus longue. Entre autres, ces patients présentent souvent une hypoxie ce qui amène une perfusion en oxygène insuffisante. Une hyperglycémie peut aussi altérer le processus de guérison [25]. Cette variable, ainsi que toutes les prochaines énumérées dans cette section sont binaires. Donc, si la caractéristique est présente, une valeur de 1 est énoncée et sinon un 0. 12% des patients avec une plaie présentent un diagnostic de diabète dans nos données.

Hypertension

En raison des mécanismes du corps humain, l'hypertension artérielle diminue l'efficacité du transport du sang, il y a donc une mauvaise circulation sanguine. Ainsi, il y a moins d'oxygène qui se rend vers la plaie pour aider la régénérescence des cellules et donc la guérison de la plaie est plus lente [12]. S'il y a présence de cette maladie chez le patient, la valeur 1 sera inscrite et 0 sinon. Seulement 11% des patients présentent ce diagnostic.

Problèmes pulmonaires

Pour des raisons similaires, c'est-à-dire par une diminution d'oxygène qui se rend vers les plaies, les problèmes pulmonaires influencent le temps de guérison d'une plaie [25]. S'il y a

présence de cette maladie chez le patient, la valeur 1 sera inscrite et 0 sinon. Moins de 2% des patients présentent un problème pulmonaire.

Obésité

De nombreuses études montrent une forte corrélation entre une augmentation du temps de guérison d'une plaie et l'obésité. Il y aurait plus de complications chez les personnes obèses. De nombreuses raisons peuvent expliquer pourquoi. Par exemple, la pression autour de leurs plaies est augmentée, ainsi que la friction autour de celles-ci [25]. La valeur 1 est inscrite si le diagnostic d'obésité est inscrit dans la base de données et 0 sinon. Moins de 1% des patients ont ce diagnostic.

Infection

S'il y a présence d'une infection au niveau de la plaie, l'inflammation sera prolongée, ce qui retardera la guérison du site. Tel qu'expliqué dans le chapitre 2, la présence de biofilms, bactéries complexes, empêche la guérison [25]. La valeur 1 est inscrite s'il y a infection et 0 sinon. Moins de 1% des patients présentent une infection à la première visite.

4.1.3 Analyse des valeurs manquantes

Un autre point important est l'analyse des valeurs manquantes. Tel qu'énoncé dans la section 3, il est important de faire une telle analyse. Le tableau 4.6 représente une liste non exhaustive des variables avec des valeurs manquantes ainsi que leur fréquence et le pourcentage sur l'ensemble des plaies. En déterminant les valeurs manquantes et en connaissant quelles sont les variables avec des valeurs manquantes, nous sommes en mesure de remplacer ces valeurs, et ce en utilisant les meilleures méthodes. Dans la section 4.2.4, des explications concernant ce que nous avons fait avec les valeurs manquantes sont énoncées.

Tableau 4.6 Analyse des valeurs manquantes

Variable	**Fréquence des valeurs manquantes**	**% l'ensemble des plaies**
Profondeur	2246	26,8 %
Âge	1666	19,9 %
Sexe	1664	19,9 %
Superficie	1037	12,4 %
Largeur	1027	12,3 %
Longueur	1013	12,1 %

4.1.4 Analyse des valeurs aberrantes et extrêmes

Dans le projet présenté, plusieurs valeurs aberrantes ont été remarquées. En effet, une plaie ayant une dimension initiale inférieure à 0.2 cm a été considérée comme une valeur aberrante. Dans la base de données, 371 plaies possèdent une superficie initiale de moins de 0.2 cm. Selon des experts en soins de plaie, il est peu probable qu'une plaie de superficie initiale de moins de 2 mm soit véridique.

Dans notre cas, les valeurs aberrantes sont souvent incluses dans les valeurs extrêmes, mais ce n'est pas toujours le cas. De plus, d'autres valeurs représentent des extrêmes. Dans le projet, les extrêmes se définissent comme les valeurs qui étaient supérieures ou inférieures à la moyenne +/- 1 écart type, puisque c'est avec cet intervalle que nos meilleurs résultats de prédiction furent trouvés.

4.1.5 Patrons de visites

En analysant la base de données, nous sommes en mesure d'explorer plusieurs aspects. Nous nous sommes intéressés à vérifier s'il y avait un patron pour la fréquence des visites. Nous remarquons donc, en analysant les graphes de la figure 4.1 qu'il n'y a pas de patron précis pour les fréquences des visites.

Les graphiques de la figure 4.1 présentent les patrons de visites pour quatre patients différents, mais ayant un profil très semblable relativement à l'âge, les diagnostics et les caractéristiques initiales de la plaie. De plus, pour ces quatre patients, il s'agissait d'une plaie de pression, mesurant initialement 1 cm^2. On remarque que chaque patient avait un patron de visite distinct, mais au bout du compte, les plaies avaient toutes environ la même durée totale de guérison, soit entre 36 et 44 jours. La dernière visite a lieu lorsque l'infirmière affirme que le patient n'a plus besoin de soins.

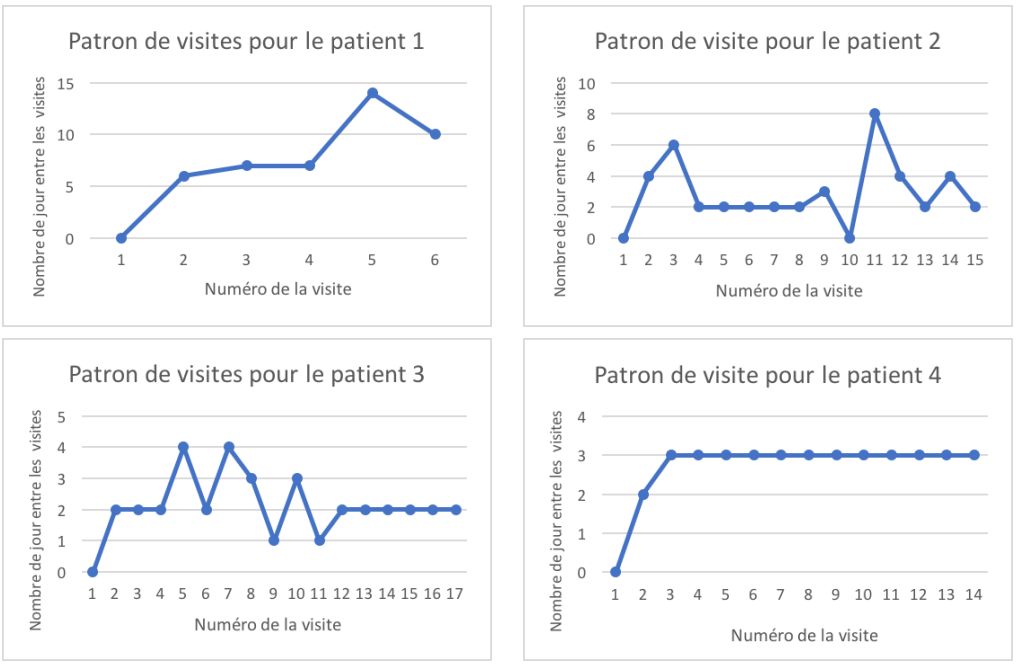

Figure 4.1 Différents patrons de visite pour différentes plaies

Par exemple, pour le patient 1, une infirmière l'a visité six fois pour faire des soins ainsi que des évaluations. L'axe des X du graphe est le numéro de la visite et l'axe des Y est le nombre de jours qui s'est écoulé entre deux visites. Toujours pour le patient 1, on remarque que pour les quatre premières visites, près d'une semaine s'était écoulée entre les visites pour ensuite monter jusqu'à deux semaines lors de la cinquième visite et puis à 10 jours pour la dernière visite. Ce patient reçut donc 6 visites sur une période de 44 jours.

Cependant, pour le patient 4, des visites beaucoup plus régulières eurent lieu. Une infirmière le visitait en moyenne aux 3 jours, et ce 14 fois sur une période de 38 jours.

Plusieurs raisons peuvent expliquer ces variations dans les patrons de visite. En effet, la présence accrue d'un proche aidant, capable de prodiguer certains soins de base pourrait faire diminuer la fréquence de visites d'une infirmière. À l'opposé, si le patient est incapable de faire ses soins et qu'il n'a pas de réseau de soutien, l'infirmière devra visiter le patient plus régulièrement. Ces informations ne sont cependant pas exactement en lien avec les plans de soins ou encore des champs spécifiques dans le logiciel d'*AlayaCare*. C'est pour ces raisons que ces informations qui nous seraient pertinentes ne sont pas connues.

4.1.6 Caractéristiques des données pour la première visite

Puisqu'un des objectifs de ce projet est d'être en mesure de prédire le temps de guérison d'une plaie à partir de la première visite d'un patient, la base de données a été modifiée pour n'avoir que les données de la première visite de chaque patient. La base de données se retrouve donc avec beaucoup moins de lignes, car en moyenne pour guérir une plaie, plus de 21 visites d'une infirmière ont lieu. La base de données modifiée contient maintenant toutes les plaies et leurs variables, mais seulement les données de la première visite pour chaque plaie sont présentes.

4.1.7 Analyse et réduction des variables explicatives

Tel qu'énoncé dans le chapitre 3, il est important de faire une analyse ainsi qu'une réduction du nombre de variables explicatives qui sera présente dans nos modèles prédictifs. Tout comme à la section 3.1.1, deux thèmes seront abordés dans cette partie du travail, soit la corrélation avec la variable cible ainsi que la corrélation entre plusieurs variables explicatives. Dans cette section, il ne sera question que de l'analyse, la réduction des variables sera faite si nécessaire dans la section 4.2.

Quelle est la corrélation avec la variable cible ?

Nous avons comparé plusieurs méthodes. Dans un premier cas, nous avions décidé de retirer chaque variable une après l'autre pour vérifier si le taux d'erreur diminuait en retirant cette variable. En utilisant le coefficient de corrélation, la variable la moins fortement corrélée avec la variable cible était retirée et ainsi de suite pour se rendre jusqu'à la plus corrélée. Si le taux d'erreur diminuait en retirant la variable, alors on retirait définitivement cette variable, si le contraire se produisait, la variable était rajoutée au modèle. Pour ce faire, nous utilisons 70% des observations, aléatoirement choisies. Pour tous les tests, nous utilisions le même échantillon.

Cependant, lorsque nous changeons l'échantillon, et que nous refaisons tout ce processus, nous retrouvons une combinaison complètement différente. Ceci nous indique que les données ne sont pas propres, ce qui signifie un risque d'y avoir beaucoup d'erreurs. Il nous est donc impossible de savoir quelle variable nous donne le plus d'informations. Toutes les variables ont donc été gardées après ce premier test.

Dans le deuxième cas, nous voulions vérifier si le coefficient de corrélation restait sensiblement le même pour chaque variable explicative si pour 10 essais, nous prenions un échantillon aléatoire.

Le tableau 4.7 présente les cinq premiers essais pour déterminer les coefficients de corrélation de certaines variables explicatives.

Tableau 4.7 Coefficients de corrélation des variables explicatives pour cinq essais

Variables/coeff	Essai 1	Essai 2	Essai 3	Essai 4	Essai 5
Acuité	0.4851	0.4636	0.1110	0.2253	0.1116
Quantité d'exsudat	0.2345	0.1596	0.2483	0.3040	0.2417
Superficie	0.1867	0.1755	0.1773	0.1640	0.1373
Odeur	-0.0864	-0.1072	-0.3426	0.0611	-0.7055
Stade	0.2875	0.0607	0.3068	0.3209	0.3036
Homme	-0.0282	0.0014	0.0179	-0.0196	0.0136
Femme	0.0282	-0.0014	-0.0179	0.0196	-0.0136
Hypertension	0.0704	0.0172	0.0698	0.1429	-0.1360
Obésité	-0.7022	-0.7712	-0.7014	-0.6562	0.0153
Diabète	-0.0748	0.0324	-0.0283	-0.3347	-0.0122

Lorsque nous analysons ce tableau, nous remarquons que les coefficients de corrélation changent parfois significativement lorsque nous utilisons un autre échantillon de données.

Lorsqu'un coefficient de corrélation change drastiquement d'un échantillon à un autre, cela signifie encore une fois qu'il y a un haut risque d'avoir des erreurs dans les données. C'est-à-dire que les informations entrées ne sont pas nécessairement véridiques. En effet, la corrélation entre les variables devrait rester sensiblement la même, peu importe l'échantillon d'observations.

L'analyse démontre qu'il n'y a pas de corrélation significative dans nos données, donc nous garderons toutes les variables pour les analyses subséquentes. Les changements dans la valeur de la variable variaient trop pour prendre une chance de retirer une certaine variable. En analysant le tableau 4.8, comprenant les valeurs des moyennes et extrêmes pour dix essais effectués avec des échantillons aléatoires, on remarque que les valeurs varient énormément. Pour avoir une forte corrélation entre deux variables, le coefficient devrait se situer entre 0,5 et 1 ou entre -0,5 et -1 [18]. Ce que nous cherchons à faire est de diminuer le nombre de variables et donc nous aurions aimé voir plusieurs variables avec une forte corrélation et ainsi rejeter celle avec une moins bonne corrélation. Nous n'avons pas été en mesure de réaliser une telle chose, car la majorité de nos variables n'avaient pas une corrélation significative avec la variable cible. À la lumière de cet essai, nous avons décidé, encore une fois, de garder l'ensemble de nos variables explicatives.

Tableau 4.8 Moyennes et extrêmes des coefficients de corrélation des variables explicatives

Variables/coeff	Moyenne essais	Valeur minimale	Valeur maximale
Acuité	0.2899	0.1110	0.4851
Quantité d'exsudat	0.2467	0.1596	0.3040
Superficie	0.1596	0.0984	0.1867
Odeur	-0.0556	-0.7055	0.6919
Stade	0.3093	0.0607	0.4246
Homme	0.0101	-0.0282	0.0383
Femme	-0.0025	-0.0179	0.0282
Hypertension	0.0074	-0.1425	0.1429
Obésité	-0.4642	-0.5854	0.0153
Diabète	-0.1569	-0.3843	0.0571

Une autre analyse fut effectuée afin de vérifier la corrélation entre les variables explicatives et la variable cible. Cette analyse consiste à évaluer le nuage de points entre la variable explicative superficie et le temps de guérison d'une plaie ainsi qu'entre l'âge et le temps de guérison d'une plaie. On peut ainsi déterminer s'il y a des tendances et le degré de corrélation.

La figure 4.2 montre un nuage de points entre une variable explicative et la variable cible, soit entre la superficie et le temps de guérison de la plaie.

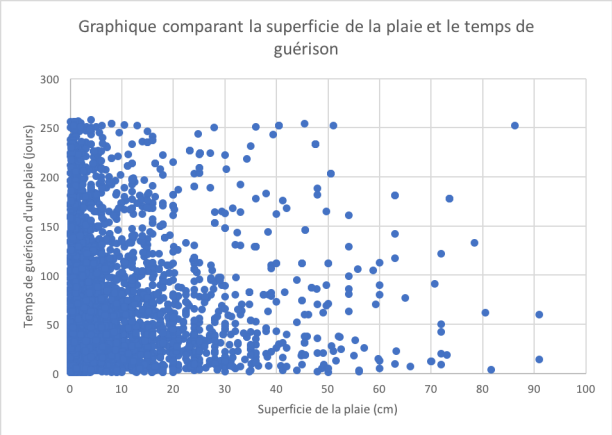

Figure 4.2 Comparaison entre la superficie de la plaie et le temps de guérison

Il parait intuitif que plus une plaie est grande, plus elle prendra de temps à guérir. Par contre, il est difficile de voir une quelconque corrélation. On remarque cependant que la majorité des observations démontrent une superficie de moins de 20 cm, cependant le temps de guérison varie énormément. De plus, nous n'observons pas la présence d'une loi normale.

Cette nouvelle analyse de la figure 4.3 consiste à vérifier la relation entre l'âge du patient et le temps de guérison.

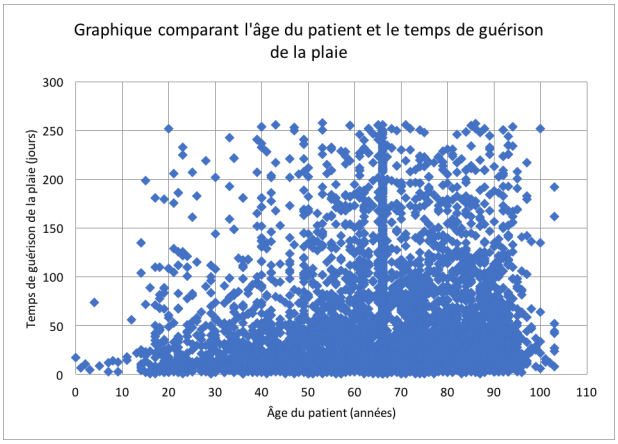

Figure 4.3 Comparaison entre l'âge du patient et le temps de guérison

On remarque qu'encore une fois, à la figure 4.3, la majorité des observations se retrouvent environ au même endroit. Il est impossible d'identifier une corrélation. De plus, pour les deux figures, il est important de comprendre que beaucoup d'autres facteurs influencent. Pour bien vérifier la corrélation, il aurait fallu avoir de nombreux cas très semblables. Présentement, on compare seulement l'âge, on ne prend pas en compte tous les autres facteurs qui influencent le temps de guérison.

En se basant sur les indices de corrélation et sur les nuages de points, il est extrêmement difficile de choisir quelles variables explicatives est nécessaire au projet, c'est pour cette raison qu'il a été décidé de tester une dernière méthode pour vérifier la pertinence de nos variables explicatives.

Pour cette méthode, il s'agit de calculer le coefficient de détermination soit le r^2. De meilleurs résultats devraient être obtenus de cette manière, de plus, comme nous avons un échantillon relativement petit, il a été décidé de calculer le r^2 sur l'ensemble de l'échantillon final disponible et non en utilisant seulement l'échantillon d'apprentissage comme ce fut le cas pour le calcul de coefficient précédent. Pour toutes nos variables explicatives, nos r^2 sont inférieurs à 0.01, donc moins de 1 % de la variabilité serait donc expliquée par la corrélation.

Quelles sont les variables explicatives fortement corrélées entre elles ?

La deuxième question qu'il faut se poser lors de l'analyse des variables explicative est la suivante : Quelles sont les variables explicatives fortement corrélées entre elles ?

La variable superficie représente la multiplication entre la largeur et la longueur. Elle est donc entièrement corrélée avec ces deux variables.

4.2 Préparation des données

Maintenant que nous connaissons bien notre base de données et tout ce qu'elle contient, il est possible de modifier efficacement nos données pour les rendre utilisables. C'est ce que nous ferons dans les prochaines sections.

4.2.1 Constitution de la base de données

Tel qu'énoncé dans les paragraphes précédents ainsi que dans l'introduction, les données recueillies se retrouvent dans trois fichiers différents. Pour les plans de soins, il s'avère plus complexe d'ajouter les informations au fichier principal, car les données sont non structurées. Ainsi, dans le but d'obtenir le plus d'informations possible sur l'état du patient et sur la plaie, du *text mining* a été réalisé dans les plans de soins. Dans les plans de soins, ce sont toutes les autres informations que l'infirmière juge pertinentes, mais où il n'y a pas de case prédéfinie à ce sujet. Nous avons donc utilisé du *text mining* pour aller vérifier s'il n'y a pas d'autres informations pertinentes qui pourraient être recueillies dans les plans de soins. Pour cela, nous avons inclus tous les synonymes en lien avec les mots cherchés, les mots au pluriel ainsi qu'au singulier, avec une majuscule ou non et avec erreur d'orthographe.

Dans un premier temps, nous nous sommes inspirés des facteurs recensés dans la littérature pour conduire notre recherche de variables pertinentes dans les plans de soins. Il a été décidé de vérifier si certains mots clés se trouvaient dans les plans de soins soit : tabagisme, tabac, cigarette, obésité, obèse, alimentation, stress, anxiété, alcool, alcoolisme, hydratation. Le tableau 4.9 résume la fréquence des mots clés que nous avons trouvés le plus fréquemment dans les plans de soins. Les mots ayant une fréquence inférieure à 400 fois ont été rejetés.

Tableau 4.9 Fréquence des nouvelles variables trouvées grâce au *text mining*

Variable	**Fréquence**
Altération	2124 fois
Détérioration	1474 fois
Infection	1103 fois
Nutrition	699 fois
Déficit	462 fois

Il y avait donc, par exemple, 1474 patients chez qui le mot détérioration se trouvait dans leur

plan de soins relatif aux plaies. On suppose que si un patient a le mot détérioration dans son plan de soins, ceci pourrait avoir une influence sur la guérison. Les fréquences données au tableau 4.9 risquent de diminuer, car il s'agit des fréquences pour l'ensemble de plans de soins et donc pas seulement pour les patients ayant une plaie.

Dans un deuxième temps, nous avons aussi vérifié quels mots étaient les plus présents dans les plans de soins. Aucune nouvelle information pertinente ne fut trouvée, c'est-à-dire aucune nouvelle variable qui influence le temps de guérison d'une plaie ne fut trouvée de cette manière.

Voici la liste des cinq mots les plus fréquents, excluant les mots de liaison :

1. Peau
2. Détérioration
3. Altération
4. Maladie
5. Image

On remarque donc que certains des mots dans cette liste ont aussi été trouvés dans l'étape précédente de *text mining*, tel que les mots détérioration et altération. Les autres mots non retenus dans cette liste n'apportaient pas d'information pertinente. En effet, les mots peau, maladie et image sont des mots trop généraux et qui ne représentent pas de facteurs qui influencent la guérison d'une plaie. De plus, nous ne retrouvions pas ces mots dans les facteurs influençant la guérison d'une plaie selon notre revue de littérature.

4.2.2 Regroupement des bases de données

Tel qu'énoncé précédemment, les données que nous souhaitions utiliser étaient dispersées dans trois bases de données différentes. Il y avait la principale, contenant toutes les informations directement liées aux plaies, c'est sur celle-ci que les autres données seront ajoutées. Premièrement, les variables énumérées au tableau 4.2 qui ont été trouvées dans les plans de soins grâce au *text mining* ont été ajoutées sous forme de nouvelles colonnes. Deuxièmement, un fichier contenant l'âge, le sexe et les diagnostics a été aussi ajouté à la base de données principale, et ce de la même manière. Il fallut aussi s'assurer que les nouvelles informations étaient inscrites pour le bon patient. Pour le sexe et l'âge, les variables ont été ajoutées telles quelles. Cependant pour les diagnostics, les maladies les plus fréquentes et celles qui pourraient avoir une influence sur la guérison d'une plaie ont été ressorties. De nouvelles colonnes binaires ont été ajoutées pour chaque diagnostic. Les autres variables ont été rejetées, car elles n'apportaient pas d'informations pertinentes à notre problématique. Le tableau 4.10 donne

les fréquences des nouvelles variables, soit les variables diagnostics ainsi que le pourcentage du temps qu'elles sont présentes dans la base de données principale.

Tableau 4.10 Fréquence des variables diagnostics

Variable	**Fréquence**	**% sur l'ensemble des plaies**
Hypertension	1691 fois	20,2 %
Diabète	908 fois	10,8 %
Problèmes pulmonaires	257 fois	3,1 %
Obésité	63 fois	0,7 %

Les fréquences présentées au tableau 4.10 concernent les données au départ, toutes les modifications n'ont pas encore été faites telles que le rejet de certaines observations. Ainsi, les fréquences peuvent donc être inférieures à la fin de toutes les modifications, car certaines valeurs seront supprimées telles qu'énoncées dans la méthodologie générale.

Les nouvelles variables, trouvées dans les plans de soins, ont donc aussi été ajoutées comme des nouvelles colonnes dans la base de données principale utilisée. La valeur 1 fut inscrite, si dans le plan de soins du patient la variable était présente et 0 si elle était absente. Ceci nous fournit une nouvelle information qui pourrait nous aider à donner de meilleures prédictions.

4.2.3 Rejet de variables

Les variables anxiété, infection, nutrition et altération ont été rejetées, car les valeurs pour ces variables étaient toutes les mêmes, soit 0. Dans notre base de données que nous utilisons pour tester nos modèles, après toutes les modifications, aucun patient ne présente ces caractéristiques. Quant à la variable stade, celle-ci fut aussi rejetée, car il y avait trop de valeur manquante, soit 6959 données manquantes sur un total de 8391 données. Il y avait donc près de 83% des valeurs qui étaient absentes pour cette variable.

4.2.4 Modification des valeurs manquantes

Pour plusieurs algorithmes, il est important de s'assurer qu'il n'y ait pas de valeur manquante sinon le modèle ne fonctionnera pas tel qu'énoncé dans le chapitre 3.

En effet, dans certains cas, il est possible de remplacer les valeurs manquantes par 0. Il en fut ainsi pour les variables suivantes : infection, déficit, anxiété, nutrition, altération, détérioration, diabète, hypertension, obésité et problèmes pulmonaires. Dans ces cas, nous

avons pris pour acquis que le patient ayant une valeur manquante pour les variables énoncées n'avait pas cette problématique. L'infirmière n'avait donc pas complété cette case ou n'avait pas écrit d'information à ce sujet dans le plan de soins pour cette raison. Cependant pour d'autres variables, il est impossible de faire ainsi, par exemple, pour l'âge ou encore pour la longueur de la plaie. Il fallait donc trouver une autre façon de gérer ces valeurs.

Intuitivement, nous trouvions que les dimensions de la plaie ainsi que certaines caractéristiques du patient ou de la plaie s'avèrent des variables trop importantes et présentant une grande variabilité pour imputer de manière simple les valeurs manquantes. Nous avons donc supprimé complètement les observations avec une valeur manquante pour plusieurs variables explicatives. [1]. Nous avons testé les trois méthodes telles que décrites dans la section 4.6. Les trois méthodes sont : supprimer les observations avec une donnée manquante, imputer par la moyenne et imputer par distribution. Nous avons gardé celle qui donnait les meilleurs résultats.

Seulement pour la variable âge, nous avons remplacé les valeurs manquantes par l'âge moyen de la base de données complète soit 66 ans. En effet, selon la littérature, l'âge n'est pas un facteur précis. Les personnes âgées de plus de 65 ans guériraient plus lentement que ceux plus jeunes, mais par la suite il n'y aurait pas de changement important. De plus, on remarque dans le tableau suivant que l'âge est manquant pour un très grand nombre de patients. Donc si nous avions supprimé toutes les observations avec l'âge manquant, nous aurions obtenu un échantillon très petit. Or, encore une fois, toutes les techniques d'imputation ont été testées.

4.2.5 Modification des valeurs aberrantes et extrêmes

Toutes les valeurs extrêmes ont été supprimées pour le modèle de régression linéaire, cependant pour les arbres et les forêts, les extrêmes ne perturbent pas le modèle. En effet, dans les modèles impliquant des arbres de décisions et les forêts aléatoires, les valeurs extrêmes sont placées dans une catégorie et n'ont pas d'influence sur le modèle. Par exemple, un patient âgé de 3 ans sera mis dans la plus petite catégorie d'âge, catégorie que le modèle aura déterminée selon le meilleur gain informationnel obtenu. La catégorie pourrait donc être, par exemple, les 30 ans et moins, le patient de 3 ans sera donc dans la même branche que celui de 30 ans. Or, pour pouvoir comparer les modèles entre eux, nous devons utiliser la même base de données. Nous avons donc supprimé les valeurs manquantes pour tous nos modèles.

Nous avons donc utilisé les équations de la section 3.2.5 pour déterminer les limites inférieures et supérieures pour chaque variable continue. Ainsi, il est possible de trouver l'intervalle où

1. Énumération des variables explicatives où nous avons supprimé les observations manquantes : emplacement, type de plaie, sexe, statut de la plaie, profondeur, largeur, longueur, acuité.

les données doivent se situer. Toutes les valeurs ne faisant pas partie de cet intervalle ont donc été supprimées.

4.2.6 Création de nouvelles variables

Parfois il est essentiel de créer de nouvelles variables pour obtenir des informations supplémentaires. Nous avons créé la variable cible Y. Ainsi, il s'agit d'une variable donnant le nombre de jours pour guérir une plaie. Pour obtenir cette valeur, nous avons soustrait la date de la dernière visite à la date de la première visite. Ainsi, nous avons obtenu un nombre de jours relatif au temps que prend la plaie à guérir. C'est donc cette variable que nous tentons de prédire.

Par la suite, plusieurs modifications de cette variable ont été réalisées. Ceci toujours dans le but de fournir les meilleurs résultats possible. Pour l'ensemble des modèles, toutes les différentes variables cibles créées ont été testées, soit le *log(y).*

La variable *log(y)* a donc été créée. Cette variable avait pour but de réduire les distances entre les données et par conséquent il y a moins d'effets des valeurs extrêmes. Il s'agit de la valeur du logarithme normal de la variable y. Les deux variables cibles présentées ci-haut sont pour les modèles de régression.

Comme nous désirions aussi évaluer si prédire un intervalle de temps au lieu de nombre précis de jours aurait de meilleurs résultats, il fallut donc créer plusieurs nouvelles variables binaires indiquant si la durée de guérison d'une plaie était dans cette catégorie.

Dans le but d'être le plus précis possible dans notre prédiction de classe, nous avons créé de nombreuses catégories, plus de 15 différentes. Nous avons fait ainsi puisque nous souhaitions obtenir une prédiction très précise. Cependant, pour plusieurs raisons hors de notre contrôle, tel qu'un nombre très restreint d'observations, il a été impossible de prédire une valeur aussi spécifique.

La figure 4.4 indique la distribution du nombre de jours pour la guérison d'une plaie selon les intervalles que nous avons choisis.

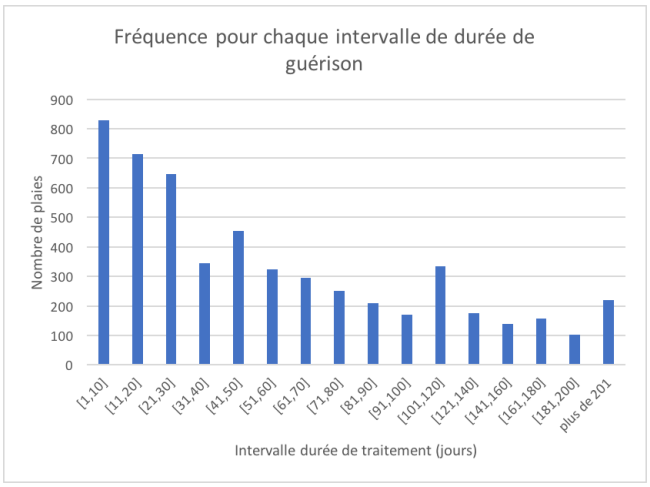

Figure 4.4 Distribution des temps de guérison d'une plaie en intervalles

On remarque qu'au départ, les intervalles présentent des bonds de 10 jours jusqu'à une durée de 100 jours, puis les intervalles augmentent pour ensuite s'arrêter à plus de 201 jours. Plusieurs raisons ont déterminé ce choix d'intervalle. Premièrement, une erreur de prédiction de 5 jours est beaucoup plus importante lorsque la durée est de 10 jours, car il s'agit d'une erreur de 50%, cependant une erreur de 5 jours pour une plaie qui prend environ 200 jours à guérir est négligeable soit 2.5%.

Une autre raison justifiant ce choix s'explique par le fait que la fréquence des plaies pour une durée plus longue est beaucoup moins importante. En effet, on s'aperçoit que plus la durée de guérison est longue, moins nous retrouvons de plaies ayant cette durée. De plus, la figure 4.4 représente l'ensemble des données et donc, puisqu'on sépare par types de plaies, les échantillons deviennent encore plus petits.

Puisqu'avec la première méthode employée pour déterminer les intervalles, nous n'avons pas été en mesure de prédire, nous avons utilisé une deuxième méthode, qui est moins précise. Le but, étant de trouver un juste milieu, c'est-à-dire des catégories incluant assez d'observations pour avoir de bonnes prédictions, mais ayant des intervalles qui apportent de nouvelles informations, alors de nouvelles catégories ont été testées. Il a été décidé de tester plusieurs autres intervalles qui selon notre jugement étaient pertinents.

Nous avons testé, sur l'ensemble des plaies, un modèle de régression logistique, en utilisant seulement deux catégories, soit si une plaie guérit entre 1 et 100 jours ainsi que l'autre catégorie, 101 jours et plus. Avec de tels intervalles, nous avions un taux de bonne classification excellent, soit de plus de 86%. Or, il n'est pas très intéressant de connaître une telle information, car une infirmière expérimentée peut certainement prédire cette valeur.

Tableau 4.11 Matrice de confusion pour la prédiction utilisant deux catégories ; test 1

Prédit/ Observé	1 à 100 jours	101 jours et plus
1 à 100 jours	1166	49
101 jours et plus	163	183

En analysant la matrice de confusion du tableau 4.11 pour ce modèle avec deux catégories, on remarque que la majorité des observations sont classées dans la catégorie de 1 à 100 jours. C'est pourquoi, nous avons décidé d'ajouter une troisième catégorie. On se retrouve donc avec trois intervalles ; 1 à 50 jours, 51 à 100 jours et finalement 101 jours et plus. En testant le même modèle que la prédiction précédente, on obtient un taux de bonne classification de près de 73%. Il s'agit d'un écart de 13% avec le modèle ayant deux catégories, cependant davantage d'informations pertinentes sont présentes dans ce cas-ci.

Tableau 4.12 Matrice de confusion pour la prédiction utilisant trois catégories ; test 2

Prédit/ Observé	1 à 50 jours	51 à 100 jours	101 jours et plus
1 à 50 jours	827	19	27
51 à 100 jours	174	62	106
101 jours et plus	62	25	257

Dans le tableau 4.12, nous observons que près de 56% des observations se retrouvent toujours dans la catégorie 0, il a été décidé de tester le modèle en utilisant encore trois catégories, mais avec des catégories différentes. Les intervalles sont de 1 à 25 jours, 26 à 50 jours puis 51 jours et plus. Avec de telles catégories, on se retrouve avec un taux de bonne classification de près de 75%.

Tableau 4.13 Matrice de confusion pour la prédiction utilisant trois catégories ; test 3

Prédit/ Observé	1 à 25 jours	26 à 50 jours	51 jours et plus
1 à 25 jours	477	12	21
26 à 50 jours	128	90	145
51 jours et plus	71	18	599

Dans cette matrice du tableau 4.13, nous remarquons qu'une vaste partie des observations se concentrent dans la dernière catégorie, c'est pourquoi le test 4 a été effectué en ajoutant une nouvelle catégorie. Un taux de bonne classification de près de 64% fut trouvé.

Tableau 4.14 Matrice de confusion pour la prédiction en utilisant quatre catégories ; test 4

Prédit/ Observé	1 à 25 jours	26 à 50 jours	51 à 100 jours	101 jours et plus
1 à 25 jours	482	14	8	6
26 à 50 jours	143	133	66	21
51 à 100 jours	67	45	121	109
101 jours et plus	17	28	42	259

En analysant la matrice de confusion du tableau 4.14, nous remarquons que près de 50% des observations se retrouvent dans la première catégorie. C'est ainsi que pour le dernier test énoncé dans ce mémoire, nous avons choisi d'ajouter une cinquième catégorie tel que présenté dans le tableau 4.15. Un taux de bonne classification de près de 55% a été trouvé avec ces intervalles.

Tableau 4.15 Matrice de confusion pour la prédiction en utilisant cinq catégories ; test 5

Prédit/ observé	1 à 12 jours	13 à 25 jours	26 à 50 jours	51 à 100 jours	101 jours et plus
1 à 12 jours	230	14	12	3	1
13 à 25 jours	110	46	64	25	5
26 à 50 jours	47	23	183	89	21
51 à 100 jours	18	5	72	138	109
101 jours et plus	1	3	37	46	259

Pour tous nos modèles de classification, nous avons donc testé en utilisant trois différents patrons d'intervalles, soit ceux que nous trouvions les plus pertinents.

Voici donc une liste des patrons d'intervalles utilisés dans notre projet :

1. 1 à 25 jours, 26 à 50 jours, 51 jours et plus
2. 1 à 25 jours, 26 à 50 jours, 51 à 100 jours, 101 jours et plus
3. 1 à 12 jours, 13 à 25 jours, 26 à 50 jours, 51 à 100 jours, 101 jours et plus

Ainsi, de nouvelles variables binaires ont été créées pour chaque intervalle. La valeur étant de 1 si le nombre de jours de guérison de la plaie est dans l'intervalle et de 0 si à l'extérieur de l'intervalle.

4.2.7 Normalisation des valeurs

Voici la liste des variables où une étape de normalisation a été faite dans notre projet :

1. Âge
2. Longueur
3. Largeur
4. Profondeur
5. Superficie

Les tableaux donnant les valeurs utilisées pour les cinq variables à normaliser ont été ajoutés à l'annexe 1. Pour chaque type de plaie, les valeurs sont différentes, car il ne s'agit pas des mêmes observations et donc il y a des moyennes et écarts types différents. L'équation utilisée pour la normalisation est énoncée dans la section de la méthodologie générale.

4.2.8 Rejet de variables explicatives

À la lumière de toutes les analyses présentées à la section 4.1.7, nous avons décidé de garder la majorité des variables explicatives pour la suite de nos expérimentations. En effet, la seule variable que nous avons rejetée est la variable superficie, car elle est la multiplication entre deux variables explicatives tel qu'énoncé dans la section 4.1.7.

4.2.9 Caractéristiques de la base de données utilisée pour les modèles

En raison des modifications nécessaires pour rendre la base de données utilisable et optimale pour les modèles de prédiction, plusieurs données ont dû être retirées dans les étapes précédentes. Ainsi, les fréquences mentionnées dans les premiers tableaux ne sont plus valables. Au départ, il y avait 8391 différentes plaies dans la base de données et au final, on en retrouve 5652. Près de 3000 observations ont donc été supprimées pour diverses raisons énoncées précédemment.

Le tableau 4.16 représente un bref récapitulatif de ce qui est maintenant présent dans la base de données.

La fréquence énoncée dans le tableau 4.16 concerne le nombre de fois que la variable binaire présente la valeur 1 et donc qui signifie que la personne possède cette caractéristique précise. Nous remarquons que la fréquence de toutes les valeurs a diminué, due au fait que plusieurs données ont été retirées. Il est intéressant de noter que toutes les variables énoncées auparavant ont gardé entre 42% et 73% de leurs données complètes et pour la majorité autour de 63%. Ainsi, les données supprimées étaient distribuées dans la majorité des variables et non seulement pour une certaine variable.

Pour les autres variables, nous remarquons que la moyenne a changé et que l'écart type

Tableau 4.16 Tableau résumé des variables binaires

Variable	Fréquence
Altération	1477 fois
Hypertension	669 fois
Diabète	709 fois
Problèmes pulmonaires	96 fois
Obésité	40 fois
Déficit	195 fois
Détérioration	943 fois
Infection	699 fois
Nutrition	445 fois

a beaucoup diminué pour la majorité des variables continues. Seulement pour la variable douleur, aucun changement n'a eu lieu. Nous pouvons expliquer cette stagnation par le fait que pour près de 64% des observations, la valeur est de 0 pour cette variable.

Tableau 4.17 Fréquence des variables explicatives continues

Variable	Moyenne	Écart type
Longueur	2.74 cm	2.44 cm
Largeur	1.72 cm	1.47 cm
Profondeur	0.52 cm	0.75 cm
Superficie	6.60 cm	10.36 cm
Douleur	1.3	2.19
Âge	66 ans	18 ans

Finalement, pour ce qui est des variables explicatives nominales, les valeurs les plus fréquentes sont restées les mêmes, car certains paramètres demeurent identiques, peu importe les modèles.

4.3 Modélisation

Tel que mentionné dans le chapitre 3, il faut déterminer quels modèles utiliser pour nos expérimentations. Par la suite, il faut utiliser des critères d'évaluation pour les comparer entre eux et ainsi déterminer quel modèle répond le mieux à nos critères d'évaluation. Il faut aussi mettre en place une marche à suivre pour tester les modèles.

Voici une liste des modèles que nous avons décidé d'utiliser pour nos tests.

1. Régression linéaire

2. Régression logistique
3. Arbre de décision de type régression
4. Arbre de décision de type classification
5. Forêt aléatoire de type régression
6. Forêt aléatoire de type classification

4.3.1 Critères d'évaluation

Les meilleurs modèles pour chaque algorithme seront comparés entre eux pour déterminer lequel sera retenu. Pour ce faire, il faut utiliser des critères d'évaluation qui sont énoncés dans les deux prochaines sous-sections.

Critères d'évaluation pour les modèles de classification

Dans notre projet, puisqu'il y a plusieurs catégories, et non une variable cible binaire, il est plus compliqué de calculer ces critères. En effet, dans notre cas, le seul critère qui concerne l'ensemble des catégories est le taux de bonne classification. Pour la sensibilité, la spécificité et la précision, il faut les calculer pour chaque intervalle. On se retrouve donc avec plusieurs valeurs pour la sensibilité, la spécificité et la précision. Il faudrait donc réaliser la matrice de confusion pour chaque catégorie distincte. Dans le but de comparer les résultats des modèles entre eux, une moyenne pour chaque critère a été calculée. Les équations pour déterminer le taux de bonne classification ainsi que les autres critères d'évaluation sont énoncés dans la section 2.2.6.

Critères d'évaluation pour les modèles de régression

Nous avons introduit des critères d'évaluation adaptés pour les algorithmes de régression. Il faut noter que nous parlons d'un nombre de jours d'erreurs moyens dans notre projet. Il ne s'agit cependant pas d'un nombre de jours réel, mais d'une erreur (calcul d'écart standard) qui pénalise particulièrement les grandes valeurs. Nous surestimons donc légèrement ce nombre de jours.

Voici la liste des critères pour les modèles de régression :

1. Nombre de jours d'erreurs moyens (EM)
2. Nombre de jours d'erreurs moyens en négatif (EN)
3. Nombre de jours d'erreurs moyens en positif (EP)

4. Fréquence d'erreurs négatives (FN)
5. Fréquence d'erreurs positives (FP)

Concernant le nombre de jours d'erreurs moyens, il s'agit de l'erreur moyenne pour toutes les plaies. Pour le critère 1, nous avons utilisé l'équation (10) ainsi que l'équation (11) pour calculer ce critère. En effet, nous utilisons la moyenne au carré pour ainsi éliminer les valeurs négatives et puis la racine carrée pour retrouver un nombre de jours d'erreurs.

$$(10) M_i = \frac{x_i - y_i}{n} \quad \forall i = 1...n$$

$$(11) EM = \sqrt{\sum M_i^2} \quad \forall i = 1...n$$

où x_i est la valeur prédite pour la plaie i et y_i est la valeur observée pour la plaie i. M est la moyenne des erreurs pour toutes les plaies. L'ensemble i regroupe toutes les plaies dans le fichier test.

Puisqu'un des objectifs cible la guérison d'une plaie le plus rapidement possible, si le modèle prédit un temps de guérison plus long que ce qui est vraiment observé, il s'agit d'une erreur moins importante. Il a donc été décidé de calculer aussi l'erreur moyenne seulement pour les valeurs négatives et seulement pour les valeurs positives.

Pour les équations (12) et (13), nous calculons l'erreur moyenne seulement pour les observations avec une erreur négative. Le calcul est par la suite le même.

Pour ce faire, les équations ci-dessous ont été utilisées :

$$(12) N_j = \frac{x_j - y_j}{n} \quad \forall j = 1...n$$

$$(13) EN = \sqrt{\sum N_j^2} \quad \forall j = 1...n$$

où x_j est la valeur prédite pour la plaie j et y_j est la valeur observée pour la plaie j. N est donc la moyenne des erreurs pour toutes les plaies. L'ensemble des plaies j représente les plaies dont l'erreur est négative dans le fichier test.

Pour les équations (14) et (15), il s'agit de calculer l'erreur moyenne où l'erreur est positive.

$$(14) P_k = \frac{x_k - y_k}{n} \quad \forall k = 1...n$$

$$(15) EN = \sqrt{\sum N_k^2} \quad \forall k = 1...n$$

où x_k est la valeur prédite pour la plaie k et y_k est la valeur observée pour la plaie i. N est donc la moyenne des erreurs pour toutes les plaies. L'ensemble des plaies k représente les plaies dont l'erreur est positive dans le fichier test.

Nous intégrons également un critère qui concerne la fréquence des erreurs positives et négatives. On peut donc savoir si régulièrement on prédit un temps de guérison plus court ou encore plus long.

Les équations (16) et (17) sont utilisées pour calculer les fréquences :

$$(16) FP = \sum x_i \geq 0 \quad \forall i = 1...n$$

$$(17) FN = \sum x_i < 0 \quad \forall i = 1...n$$

où x_i est la valeur prédite pour la plaie i. L'ensemble i représente toutes les plaies dans le fichier test.

Pour ce modèle, nous avons tenté de prédire seulement la variable y, mais en utilisant plusieurs combinaisons possibles pour les intervalles.

Le nombre de catégories pour la variable cible est choisi par l'analyste. Dans notre cas, nous avons testé de nombreuses combinaisons possibles de catégories. Cela représentait tout un défi de trouver le bon nombre de catégories, car moins il y en a et plus les intervalles sont larges et donc moins précis.

4.3.2 Techniques pour vérifier la présence de sur-apprentissage

Tel que précisé dans la section 4.3.2, il est primordial d'utiliser des moyens pour déterminer la présence de sur-apprentissage. Plusieurs techniques s'offrent à nous, tels que l'utilisation de deux fichiers distincts et la validation croisée.

Dans le projet, nous avons opté pour ces deux techniques. Nous avons testé tous nos modèles avec la technique des deux fichiers distincts, car la majorité du temps nous obtenions de meilleurs résultats en utilisant cette technique. Les résultats des tests effectués entre les deux techniques pour éviter le sur-apprentissage sont présentés dans la section 5.1.5.

Avec cette technique, nous devions fixer un écart maximal entre nos deux fichiers, soit celui d'apprentissage et celui de test. En raison de la grosseur de nos échantillons, nous acceptions un écart maximal de 6% entre le fichier d'apprentissage et le fichier test. Une différence supérieure à ce pourcentage signifierait que notre modèle présente du sur-apprentissage.

Pour les modèles de régression, nous acceptions un erreur maximal de 6 jours entre les deux fichiers. Avec une différence inférieur à 6 jours, nous prenions pour acquis qu'il n'y avait pas de sur-apprentissage.

4.3.3 Récapitulatif de la base de données utilisée à tous les tests

Nous avons testé tous les modèles en utilisant la même base de données. Le tableau 4.18 représente toutes les variables explicatives présentes dans la base de données pour tous les tests effectués avec les modèles prédictifs.

Tableau 4.18 Variables explicatives présentes dans la base de données

Caractéristiques du patient	**Caractéristiques de la plaie**
Diabète	Longueur, largeur, profondeur
Âge	Type de plaie
Sexe	Statut de la plaie
Hypertension	Espace sous-jacent
Problèmes pulmonaires	Tunnel
Détérioration	Odeur
Déficit	Acuité

Il est important de noter que pour la variable type de plaie, à certains moments, cette variable explicative était rejetée. En effet, tel qu'affirmé à la section 4.1.2, nous avons séparé notre base de données par type de plaie. Donc, c'est seulement au moment où nous testons notre modèle prédictif sur l'ensemble des plaies que cette variable explicative est présente.

Nous avons donc quatre bases de données, contenant toutes les mêmes variables, sauf la variable type de plaie pour les trois dernières bases de données. Le nombre d'observations dans chaque base de données est différent.

1. Base de données contenant l'ensemble des plaies
2. Base de données contenant seulement les plaies traumatiques
3. Base de données contenant seulement les plaies opératoires
4. Base de données contenant seulement les plaies de pression

Tel qu'énoncé dans la section 4.2.6, nous avons créé des groupes pour les variables cibles pour faire différents tests dans les modèles de classification. Le tableau 4.19 représente les différents groupes. Ces variables deviennent donc nos variables cibles pour les modèles de classification. Par exemple, lors de l'utilisation du groupe 1, le modèle tentera de prédire dans lequel des trois intervalles l'observation est présente.

Tableau 4.19 Différents groupes pour les modèles de classification

Groupe	Intervalles présents
Groupe 1	1 - 25 jours, 26 - 50 jours et de 51 jours et plus.
Groupe 2	1 - 25 jours, 26 - 50 jours, 51 - 100 jours et 101 et plus.
Groupe 3	1 - 12 jours, 13 - 25 jours, 26 - 50 jours, 51 - 100 jours et 101 et plus.

Toujours dans la section 4.2.6, nous vous avons présenté nos variables cibles pour les modèles de régression. La liste suivante est donc une énumération de ces variables.

1. Y, soit le temps entre la première visite et la dernière visite
2. Log(Y)

4.3.4 Paramètres généraux pour l'expérimentation

Tous les résultats ont été obtenus en utilisant le logiciel *PyCharm*, version 2016.1.2. Pour les modèles étudiés, nous avons utilisé la librairie *sk.learn* qui contient tous les modèles.

Voici la liste de tous les modèles utilisés :

— *LinearRegression*
— *LogisticRegression*
— *DecisionTreeClassifier*
— *DecisionTreeRegressor*
— *RandomForestClassifier*
— *RandomForestRegressor*

De plus, à plusieurs endroits dans ce chapitre, nous ferons état du *random state*. Le *random state* est un paramètre que nous avons utilisé dans tous les modèles. Il sert à toujours choisir le même échantillon lorsque nous testons le modèle. En effet, si nous posons ce paramètre égal à 5, par exemple, le modèle prendra toujours le même patron d'observations. Nous pourrons donc comparer par la suite les modèles entre eux, car le même échantillon aura été sélectionné. Si nous ne posons pas un *random state*, le modèle prendra aléatoirement à chaque essai les observations de la base de données.

4.3.5 Paramètres pour la régression linéaire

Pour la régression linéaire, nous avons utilisé le modèle *LinearRegression*. Les paramètres suivants étaient présents dans le modèle : *fit.intercept=True, normalize=False,copy.X=True, n.jobs=1 et random.state=5.* Nous les avons tous laissé par défaut, excepté le *random state*

qui est égal à 5 pour tous les essais.

4.3.6 Paramètres pour la régression logistique

Pour la régression logistique, nous avons utilisé le modèle *LogisticRegression*. Les paramètres suivants étaient présents dans le modèle : *penalty='l2', dual=False, tol=0.0001,C=1.0, fit.intercept=True, intercept.scaling=1, class.weight=None, random.state=5, solver='liblinear', max.iter=100, multi.class='ovr', verbose=0, warm.start=False et n.jobs=1.* Tout comme la régression linéaire, nous avons laissé tous les paramètres par défaut à l'exception du *random state*.

4.3.7 Paramètres pour les arbres de décision de type classification

Pour les arbres de décision de type classification, nous avons utilisé le modèle *DecisionTreeClassifier*. Les paramètres suivants étaient présents dans le modèle : *criterion='gini', splitter='best', max.depth=None, min.samples.split=2, min.samples.leaf=1, min.weight=0.0, max.features=None, random.state=5, max.leaf=None, class.weight=None, presort=False.*

Pour ce modèle, nous avons modifié deux paramètres soit le *max.depth* et le *criterion*. Le premier paramètre que nous avons modifié est la profondeur de l'arbre. Pour ce critère, nous avons testé les entiers entre 1 et 10. Le second paramètre concerne le critère de séparation qu'utilisera le modèle pour séparer les branches de son arbre. Deux valeurs sont possibles, il s'agit du critère d'Entropie et le critère de Gini. Nous avons donc testé 20 combinaisons possibles entre nos deux paramètres. La combinaison nous donnant les meilleurs résultats, et ce en évitant le sur-apprentissage était enregistrée pour chaque type de plaie.

4.3.8 Paramètres pour les arbres de décision de type régression

Pour les arbres de décision de type régression, nous avons utilisé le modèle *DecisionTreeRegressor*. Les paramètres suivants étaient présents dans le modèle : *criterion='mse', splitter='best', max.depth=None, min.samples.split=2, min.samples.leaf=1, max.features=None, random.state=5, max.leaf.nodes=None, min.impurity.split=None et presort=False.*

Pour ce modèle, nous avons seulement modifié le paramètre profondeur. Encore une fois, nous avons testé toutes les valeurs de 1 à 10 pour ce paramètre pour trouver le meilleur résultat, mais en s'assurant d'éliminer le sur-apprentissage.

4.3.9 Paramètres pour les forêts aléatoires de type classification

Pour les forêts aléatoires de type classification, nous avons utilisé le modèle *RandomForestClassifier*. Les paramètres suivants étaient présents dans le modèle : *n.estimators=10, criterion='gini', max.depth=None, min.samples.split=2, min.samples.leaf=1, max.features='auto', max.leaf.nodes=None, min.impurity.decrease=0.0, min.impurity.split=None, bootstrap=True, oob.score=False, n.jobs=1, random.state=5, verbose=0, warm.start=False, class.weight=None*

Pour ce type de modèle, nous avons modifié trois paramètres soit *n.estimators* qui signifie le nombre d'arbres présents dans la forêt, le *criterion* et *max.depth*. Pour le paramètre du nombre d'arbres, nous avons testé de $n = 1$ jusqu'à 10. Les deux autres paramètres peuvent avoir les mêmes valeurs que dans les arbres de décision. Nous avons testé toutes les combinaisons possibles, et ce pour chaque base de données et nous avons gardé seulement la meilleure combinaison.

4.3.10 Paramètres pour les forêts aléatoires de type régression

Pour les forêts aléatoires de type régression, nous avons utilisé le modèle *RandomForestRegressor*. Les paramètres suivants étaient présents dans le modèle : *n.estimators=4, criterion='mse', max.depth=None, min.samples.split=2, min.samples.leaf=1, max.features='auto', max.leaf.nodes=None, min.impurity.decrease=0.0, min.impurity.split=None, bootstrap=True, oob.score=False, n.jobs=1, random.state=5, verbose=0, warm.start=False.*

Dans ce modèle, nous avons modifié et testé de nombreuses combinaisons pour deux paramètres soit le *n.estimators* et le *max.depth*. La meilleure combinaison pour chaque base de données était enregistrée. Nous avons, encore une fois, testé tous les entiers de 1 à 10 pour ces deux critères.

.

CHAPITRE 5 RÉSULTATS ET DISCUSSION

Ce chapitre est séparé en deux sections. Premièrement, de nombreux tableaux présentent les résultats obtenus pour chaque type de plaie avec chaque modèle. L'autre partie résume l'analyse et la discussion en lien avec les résultats obtenus.

5.1 Résultats obtenus

5.1.1 Résultats pour l'ensemble des plaies

Le premier test consiste à faire une analyse pour tous les types de plaies confondus. Nous désirions vérifier si de meilleurs résultats pouvaient être obtenus de cette manière. En faisant ainsi, nous avions un échantillon plus grand pour tester les modèles soit de 3311 observations. La diminution de la taille de l'échantillon est causée par toutes les modifications apportées à la base de données telle que mentionnée dans le chapitre 4.

De plus, pour tous les tests effectués dans ce chapitre, les observations étaient divisées en deux fichiers. Le premier, le fichier test contenant 30% des observations et le fichier d'apprentissage qui contenait 70%.

Modèles de régression

Le tableau 5.1 donne les résultats pour les modèles de type régression pour un échantillon contenant l'ensemble des plaies. Nous comparons les résultats obtenus en fonction des critères de régression dans ce cas-ci. Nous avons donc testé l'échantillon avec les modèles de régression linéaire, d'arbre de décision et de forêts aléatoires. Il n'y a pas présence de sur-apprentissage, car on remarque que la différence entre les résultats du fichier d'apprentissage et celui test n'est pas significative. Tel que mentionné à la section 4.3.2, nous acceptions une différence inférieure à 6%.

Tel qu'énoncé dans le chapitre précédent, dans le but d'avoir le même échantillon, le *random state* avait la valeur 5 pour tous les modèles testés.

Le seul paramètre qui fut changé, pour l'arbre de régression, concerne la profondeur maximale, la valeur qui donnait les meilleurs résultats est 5 couches maximum. Si nous augmentions la profondeur, de meilleurs résultats étaient obtenus, or il y avait présence de sur- apprentissage.

Pour les forêts de type régression, la profondeur maximale inscrite fut 7 couches et le nombre d'arbres que contient la forêt était de 10 arbres.

Tableau 5.1 Résultats pour l'ensemble des plaies avec modèles régressions

Modèles utilisés	Erreurs moyens fichier test	Erreurs moyens fichier train	Erreurs moyens négatifs	Erreurs moyens positifs	Fréquence d'erreurs négatives	Fréquence d'erreurs positives
Régression linéaire	38.90 jours	39.22 jours	36.87 jours	20.28 jours	376 fois	618 fois
Arbre de décision régression	36.70 jours	35.34 jours	30.52 jours	20.43 jours	411 fois	582 fois
Forêt aléatoire régression	32.78 jours	28.92 jours	27.83 jours	18.12 jours	399 fois	595 fois

On remarque que le meilleur modèle pour prédire l'ensemble des plaies, en utilisant un algorithme de régression, est la forêt aléatoire. L'erreur moyenne pour ce modèle est de plus de 32 jours, comparativement à près de 37 jours ainsi que près de 39 jours.

Modèles de classification

Voici les résultats pour les modèles de type classification pour l'échantillon contenant l'ensemble des plaies. Pour tous les essais, le *random state* avait une valeur de 5. Pour les arbres de décision, les meilleurs résultats étaient obtenus en utilisant une profondeur maximale de 4 avec comme critère de séparation *Gini*. De plus, les meilleurs résultats pour les forêts aléatoires étaient obtenus avec les mêmes valeurs pour ces deux critères. Le nombre d'arbres maximal pour la forêt était de 7, et ce pour tous les intervalles.

Tableau 5.2 Résultats ensemble plaies ; avec modèles classification

Modèles	Taux bonne classification test	Taux bonne classification train	Sensibilité moyenne	Spécificité moyenne	Précision moyenne
Groupe 1					
Régression logistique	71,22%	76,62%	67,87%	84,98%	71%
Arbre décision classification	71,13%	75,78%	69,67%	85,57%	71%
Forêt aléatoire classification	70,12%	75,27%	66%	83,91%	71%
Groupe 2					
Régression logistique	58,65%	63,88%	58,41%	86,12%	58%
Arbre décision classification	58,35%	63,53%	57,12%	85,96%	62%
Forêt aléatoire classification	54,02%	60,68%	52,88%	84,63%	55%
Groupe 3					
Régression logistique	51,31%	56,32%	49,52%	87,43%	49,20%
Arbre décision classification	52,52%	57,27%	53%	87,86%	57%
Forêt aléatoire classification	47,10%	53,17%	47,34%	86,41%	48%

Dans le tableau 5.2, nous observons que le meilleur taux de bonne classification est de plus de 71%, et ce avec le groupe 1. Tous les groupes dont il est question dans ce chapitre sont les groupes énumérés dans la section 4.3.3.

5.1.2 Résultats plaies opératoires

Maintenant pour l'échantillon de données contenant seulement les observations appartenant à une plaie opératoire. Nous retrouvons 555 observations dans cet échantillon.

Modèle de régression

Le tableau 5.3 énumère les résultats des meilleurs modèles de régression pour les plaies opératoires.

Dans ce cas, pour l'arbre de décision, la profondeur maximale était de 3. De plus, avec le modèle de forêt aléatoire, une profondeur maximale de 4 fut utilisée et il y avait 4 arbres

dans la forêt. Le taux d'erreurs moyen le plus petit obtenu avec nos tests est de 20.05 jours d'erreur.

Tableau 5.3 Résultats plaies opératoires avec modèles régressions

Modèles utilisés	Erreurs moyens fichier test	Erreurs moyens fichier train	Erreurs moyens négatifs	Erreurs moyens positifs	Fréquence d'erreurs négatives	Fréquence d'erreurs positives
Régression linéaire	22.13 jours	21.60 jours	17.04 jours	13.21 jours	88 fois	153 fois
Arbre de décision régression	20.52 jours	19.92 jours	18.81 jours	11.73 jours	71 fois	129 fois
Forêt aléatoire régression	20.05 jours	14.68 jours	16.65 jours	12.84 jours	75 fois	126 fois

Modèles de classification

Le tableau 5.4 expose les résultats pour les plaies opératoires en utilisant des modèles de classification.

Pour le groupe 1, pour le modèle d'arbre et de la forêt, le critère d'*Entropie* a été utilisé. De plus, ces modèles avaient une profondeur maximale de 2. La forêt contenait 4 arbres.

Pour le second groupe, il s'agit encore du critère de séparation *Entropie* qui fut utilisé pour les deux modèles nécessitant un critère de séparation, tous deux présentaient une profondeur maximale de 4. Pour la forêt, 4 arbres furent nécessaires pour obtenir le meilleur résultat.

Pour le troisième groupe, dans le modèle de l'arbre de décision, le critère d'entropie fut utilisé et il avait une profondeur maximale de 3. Quant à la forêt aléatoire, encore le critère *Entropie*, mais une profondeur maximale de 4 et il y avait 9 arbres pour la forêt.

Avec cet échantillon, nous obtenons presque les mêmes résultats avec le groupe 1 ainsi que le groupe 2, soit un taux de bonne classification maximale de près de 69%.

Tableau 5.4 Résultats plaies opératoires : avec modèles classification

Modèles	Taux bonne classification test	Taux bonne classification train	Sensibilité moyenne	Spécificité moyenne	Précision moyenne
Groupe 1					
Régression logistique	68,66%	73,34%	67,74%	83,87%	68%
Arbre décision classification	68,66%	71,43%	65,43%	83,96%	73%
Forêt aléatoire classification	64,18%	69,30%	63,64%	82,23%	66%
Groupe 2					
Régression logistique	64,18%	70,15%	59,13%	86,40%	63%
Arbre décision classification	68,16%	73,13%	62,66%	88%	68%
Forêt aléatoire classification	68,66%	72,49%	53,91%	87,98%	68%
Groupe 3					
Régression logistique	56,72%	60,13%	53,11%	88,11%	54%
Arbre décision classification	59,70%	59,91%	65,12%	89,54%	59%
Forêt aléatoire classification	56,72%	61,83%	49,60%	88%	58%

5.1.3 Résultats plaies traumatiques

Maintenant, pour l'échantillon des plaies traumatiques où l'échantillon contenant 583 observations. De nombreux essais furent à nouveau tentés avec les observations pour les plaies traumatiques.

Modèle de régression

Premièrement, le tableau 5.5 donne les résultats pour les modèles de type régression.

Pour l'arbre de décision, le critère d'*Entropie* fut utilisé. La profondeur maximale était de 3 couches. De plus, avec le modèle de forêt aléatoire, cette fois-ci, le meilleur résultat était obtenu avec le critère de *Gini*, une profondeur maximale de 3 fut appliquée et il y avait 4 arbres dans la forêt. Une erreur moyenne de plus de 23 jours est obtenue avec le modèle de forêt aléatoire.

Tableau 5.5 Résultats plaies traumatiques avec modèles régressions

Modèles utilisés	Erreurs moyens fichier test	Erreurs moyens fichier train	Erreurs moyens négatifs	Erreurs moyens positifs	Fréquence d'erreurs négatives	Fréquence d'erreurs positives
Régression linéaire	24.97 jours	21.18 jours	22.46 jours	14 jours	66 fois	80 fois
Arbre de décision régression	25.26 jours	26.34 jours	23.42 jours	14 jours	63 fois	83 fois
Forêt aléatoire régression	23.42 jours	25.73 jours	21.45 jours	11.71 jours	65 fois	81 fois

Modèle de classification

Le tableau 5.6 est les résultats pour chaque groupe pour les modèles de classification pour les plaies traumatiques.

Pour ce premier groupe, *Gini* est le critère de séparation qui fut utilisé pour l'arbre et la forêt. Pour le modèle utilisant l'arbre de décision, la profondeur maximale était de 5. Concernant la forêt aléatoire, la profondeur maximale était de 6 et il y avait 7 arbres pour faire ce modèle.

Pour le deuxième groupe, le critère d'*Entropie* fut utilisé pour l'arbre et la forêt. L'arbre de décision avait une profondeur maximale de 3. Pour la forêt aléatoire, une profondeur de 4 nous donnait les meilleurs résultats et elle contenait 2 arbres.

Finalement pour le dernier groupe de classification pour les plaies traumatiques, le critère de *Gini* fut utilisé avec une profondeur maximale de 3 autant pour l'arbre que pour la forêt. De plus, pour la forêt aléatoire, il y avait 3 arbres.

Pour les plaies traumatiques, avec la régression logistique, pour le groupe 1, nous obtenons un taux de bonne classification de près de 80%, ce qui représente le meilleur résultat que nous avons obtenu avec cet échantillon.

Tableau 5.6 Résultats plaies traumatiques : avec modèles classification

Modèles	Taux bonne classification test	Taux bonne classification train	Sensibilité moyenne	Spécificité moyenne	Précision moyenne
Groupe 1					
Régression logistique	79,45%	73,76%	75,92%	89,44%	79%
Arbre décision classification	76,71%	80,83%	74,61%	88,65%	78%
Forêt aléatoire classification	76,03%	81,41%	75,05%	88,50%	78%
Groupe 2					
Régression logistique	68,49%	69,49%	62,79%	89,04%	68%
Arbre décision classification	64,38%	67,25%	56,99%	87,88%	66%
Forêt aléatoire classification	64,38%	67,85%	58,67%	87,61%	66%
Groupe 3					
Régression logistique	56,85%	61,06%	50,66%	88,85%	52%
Arbre décision classification	53,42%	60,18%	50,97%	71,04%	57%
Forêt aléatoire classification	50%	58,11%	42,96%	87,10%	37%

5.1.4 Résultats plaies de pression

Voici la dernière section des résultats, cette fois-ci pour l'ensemble de plaies de pression. Cet échantillon est celui qui contient le moins d'observations soit 420, et ce suite à toutes les suppressions d'observations.

Modèle de régression

Les résultats pour les modèles de régression sont présentés au tableau 5.7 pour les plaies de pression.

La profondeur maximale était de 4 couches. De plus, avec le modèle de forêt aléatoire, cette fois-ci le meilleur résultat était obtenu avec une profondeur maximale de 5 et il y avait 8 arbres dans la forêt.

Encore une fois, le meilleur résultat obtenu pour un modèle de régression est la forêt aléatoire avec un taux d'erreur de près de 24 jours pour les plaies de pression.

Tableau 5.7 Résultats plaies de pression avec modèles régressions

Modèles utilisés	Erreurs moyens fichier test	Erreurs moyens fichier train	Erreurs moyens négatifs	Erreurs moyens positifs	Fréquence d'erreurs négatives	Fréquence d'erreurs positives
Régression linéaire	33,57 jours	34.86 jours	26,1 jours	23,33 jours	58 fois	60 fois
Arbre de décision régression	28,50 jours	23,78 jours	21,42 jours	19,23 jours	49 fois	69 fois
Forêt aléatoire régression	23,91 jours	21,57 jours	15,50 jours	17,25 jours	55 fois	63 fois

Modèles de classification

Le tableau 5.8 représente les résultats pour les plaies de pression en utilisant les modèles de classification.

Pour le groupe 1 des plaies de pression, le critère de *Gini* fut utilisé pour les modèles d'arbre et de forêt. Pour l'arbre de décision, une profondeur maximale de 5 nous donnait d'excellents résultats. Pour la forêt, une profondeur de 6 était encore mieux avec 6 arbres pour composer la forêt.

Pour le deuxième groupe de ce type de plaie, nous avons encore utilisé le critère de *Gini*. Cette fois le modèle de l'arbre de décision avait une profondeur maximale de 5 et celui de la forêt de 4. Pour la forêt, 5 arbres ont été employés.

Finalement, il s'agit encore du critère de *Gini* qui fut utilisé pour ce troisième groupe. Pour l'arbre de décision, une profondeur maximale de 3 a été utilisée et pour la forêt la valeur 4 fut employée pour la profondeur. Trois arbres furent utilisés pour la forêt aléatoire dans ce cas.

Nous obtenons un taux de bonne classification de plus de 86% en utilisant la forêt aléatoire pour le groupe 1. Il s'agit du meilleur résultat pour le tableau 5.8.

Tableau 5.8 Résultats plaies de pression : avec modèles de classification

Modèles	Taux bonne classification test	Taux bonne classification train	Sensibilité moyenne	Spécificité moyenne	Précision moyenne
Groupe 1					
Régression logistique	77,97%	76,73%	75,75%	87,80%	78%
Arbre de décision classification	81,36%	86,91%	80,25%	90,21%	82%
Forêt aléatoire classification	86,44%	88,73%	81,80%	91,84%	86%
Groupe 2					
Régression logistique	63,56%	66,55%	65,12%	87,95%	66%
Arbre de décision classification	73,73%	77,82%	71,15%	90,67%	74%
Forêt aléatoire classification	72,03%	76,36%	71,19%	90,37%	72%
Groupe 3					
Régression logistique	61,02%	65,09%	53,27%	88,38%	59%
Arbre de décision classification	62,95%	66,18%	56,59%	90,53%	74%
Forêt aléatoire classification	63,56%	66,91%	55,38%	90,73%	63%

5.1.5 Modèle avec utilisation de validation croisée

L'utilisation de deux fichiers, soit d'apprentissage et de test, nous a donné des résultats acceptables, cependant comme la base complète de données ne contient pas un nombre astronomique d'observations. Tel que noté précédemment, une autre option est possible pour empêcher le sur-apprentissage. Il s'agit de la validation croisée.

Dans notre projet, pour le nombre de groupes dans la validation croisée, nous avons utilisé n = 5, puisque notre échantillon est relativement petit. Il a été décidé de tester seulement les modèles de régression pour chaque plaie et aussi le modèle de classification régression logistique du groupe 1, soit avec trois intervalles. Nous désirions vérifier si l'utilisation de cette technique améliorerait nos résultats.

Le tableau 5.9 fournit un comparatif des taux de bonne classification pour le groupe 1 de chaque plaie en utilisant soit deux fichiers distincts ou la validation croisée.

Tableau 5.9 Tableau comparatif des taux de bonnes classifications

Type de plaie	**2 fichiers différents**	**Validation croisée**	**Différence entre les deux techniques**
Ensemble plaies	71,22 %	73,80 %	2,58
Plaies opératoires	68,66 %	70,30 %	1,64
Plaies traumatiques	79,45 %	72,20 %	-7,25
Plaies de pression	77,97 %	72,50 %	-5,47

Nous remarquons que la moitié du temps, pour les modèles de classification, il y a une meilleure prédiction d'environ 2% avec la validation croisée, cependant nous perdons entre 5 et 7% de taux de bonne prédiction pour les deux autres ensembles.

Le tableau 5.10 représente le nombre de jours d'erreurs moyens pour chaque plaie en utilisant soit deux fichiers distincts ou la validation croisée pour les modèles de régression.

Tableau 5.10 Tableau comparatif du nombre de jours d'erreurs

Type de plaie	**2 fichiers différents**	**Validation croisée**	**Différence entre les deux techniques**
Ensemble plaies	38,90 jours	52,46 jours	13,56 jours
Plaies opératoires	22,13 jours	22,87 jours	0,74 jour
Plaies traumatiques	24,97 jours	28,03 jours	3,06 jours
Plaies de pression	33,57 jours	37,13 jours	3,56 jours

Pour les modèles de régression, dans tous les cas, les prédictions utilisant la technique des deux fichiers distincts, nous permettent d'obtenir de meilleurs résultats. Nous avons donc privilégié l'utilisation des fichiers d'apprentissage ainsi que de test et non de la validation croisée pour tous nos tests.

5.1.6 Résumé des meilleurs résultats obtenus

Le tableau 5.11 résume les meilleurs résultats obtenus pour chaque ensemble de données.

Il est impossible de déterminer entre les modèles de régression et de classification lequel est le

meilleur car ceci dépend toujours de l'objectif. Pour certains gestionnaires, il serait préférable pour eux d'obtenir une prédiction avec seulement trois intervalles, mais pour d'autres ils préféreraient plus d'intervalles, en acceptant un taux d'erreurs plus grand.

De plus, il est difficile d'affirmer quel modèle donne les meilleurs résultats. Prenons l'exemple des plaies opératoires, le tableau 5.11 expose une comparaison des meilleurs modèles pour chaque intervalle en utilisant seulement le taux de bonne classification du fichier test.

Tableau 5.11 Tableau des meilleurs résultats pour chaque ensemble de données

Type de plaies	Nombre de jours d'erreurs	Meilleur modèle régression	Taux de bonne classification	Meilleur modèle classification
Ensemble plaies	32,78 jours	Forêt aléatoire	71,22%	Régression logistique
Plaies opératoires	20,05 jours	Forêt aléatoire	68,66%	Forêt aléatoire
Plaies traumatiques	23,42 jours	Forêt aléatoire	79,45%	Régression logistique
Plaies de pression	23,91 jours	Forêt aléatoire	86,44%	Forêt aléatoire

Dans cet exemple du tableau 5.12, on peut facilement rejeter le groupe 1, puisque le taux de bonne classification est le même que pour le second groupe et qu'on sait que le groupe 1 est moins précis pour le nombre de jours. Cependant pour les deux autres groupes, il est difficile de déterminer le meilleur, car chacun apporte un avantage sur l'autre.

Tableau 5.12 Tableau comparaison des meilleurs résultats par classification pour les plaies opératoires

Groupe d'intervalle	Taux de bonne classification
Groupe 1	68,66%
Groupe 2	68,66%
Groupe 3	59,70%

Un autre exemple avec lequel il est encore plus difficile de choisir le meilleur modèle concerne les plaies de pression du tableau 5.13.

Dans ce cas, nous constatons bien que plus on compte d'intervalles différents, moins le taux de bonne classification est élevé. Il n'y a pas de meilleur modèle précis, cela dépend toujours de l'objectif principal.

Tel qu'indiqué dans la méthodologie ainsi que dans les tableaux de résultats et dans la littérature, des critères d'évaluations plus spécifiques sont disponibles. Les gestionnaires pourront

Tableau 5.13 Tableau comparaison des meilleurs résultats par classification pour les plaies de pression

Groupe d'intervalle	Taux de bonne classification
Groupe 1	86,44%
Groupe 2	73,73%
Groupe 3	63,56%

donc choisir le modèle qui leur semble le plus pertinent en utilisant aussi la spécificité ou encore la sensibilité.

5.1.7 Qualité de nos résultats

D'un point de vue mathématique, nos résultats nous permettent de prédire et d'obtenir des balises pertinentes à la prédiction du temps de guérison des plaies. Par contre, pour un gestionnaire de ressources infirmières, ces résultats finaux ne s'avèrent utiles que dans certains cas, c'est ce dont il sera question dans les prochains paragraphes de cette section.

Concernant la qualité de nos résultats, nous proposons trois façons d'en discuter. La première consiste à faire une comparaison par rapport à la littérature pour déterminer si le taux de prédiction produit par nos modèle est de bonne qualité. La seconde consiste à déterminer si l'erreur (en nombre de jours) obtenue est raisonnable. Finalement, la troisième façon consiste à déterminer si les résultats sont utilisables en pratique, par exemple par les gestionnaires.

Selon la première analyse, il est très difficile de comparer nos résultats avec ceux de la littérature, et ce, pour plusieurs raisons. Premièrement, dans la revue de littérature, nous n'avions pas été en mesure de trouver des articles de prédiction du temps de guérison d'une plaie, donc nous ne pouvons pas affirmer si nos modèles sont meilleurs ou pires. Deuxièmement, la majorité des articles que nous avons trouvés sur la prédiction d'événements en santé tentaient de prédire une variable binaire et non à plusieurs intervalles. Il est donc encore une fois difficile de comparer nos résultats avec ceux-là. Tel qu'énoncé dans le chapitre 2, pour les articles que nous avons trouvés, les taux de bonnes classifications variaient entre 75% et 100%, et ce pour une prédiction de variables cibles binaires. On peut donc affirmer que dans certains cas, nos résultats sont semblables à ceux de la littérature, car pour l'ensemble des plaies de pression, pour le groupe 1, nous avons un taux de bonnes classifications de 86,44%. Avec d'autres modèles, nos résultats sont inférieurs aux moyennes présentées dans la revue de littérature. Il faut cependant garder en tête que dans nos résultats, nous utilisons plus de deux intervalles de prédiction.

Pour la deuxième analyse, soit en lien avec le nombre de jours d'erreurs obtenus. Il est encore plus ardu d'affirmer ou d'infirmer si nos résultats sont raisonnables, lorsque nous les comparons à la littérature. Nous n'avons pas été en mesure de trouver des articles traitant de sujets similaires où nous aurions pu comparer nos résultats. Plusieurs facteurs doivent être pris en compte pour les erreurs de régression. En effet, à partir de nos modèles, nous sommes parvenus à une moyenne d'erreurs de 27 jours et la moyenne des soins reçus par nos patients était de plus de 64 jours. Il est important de vérifier le nombre de jours total, car une erreur de 27 jours sur 64 jours n'est pas la même chose qu'une erreur de 27 jours sur 30 jours par exemple. À ce stade, nous ne pouvons donc pas affirmer si nos résultats pour les modèles de régression sont satisfaisants ou non, car nous n'avons pas été en mesure de trouver des comparatifs. Dans le prochain paragraphe, nous élaborerons comment le nombre de jours d'erreurs peut jouer dans la pratique.

Pour la dernière analyse concernant la qualité de nos résultats, c'est-à-dire vérifier si nos résultats sont utilisables en pratique. Nous croyons que ceux-ci peuvent être utiles dans certaines situations. Prenons le cas d'un gestionnaire d'agence d'infirmières qui construit les horaires à chaque mois, donc sur un horizon de temps de 31 jours. En utilisant les prédictions que notre modèle est en mesure de faire, il pourra confectionner son horaire en ayant un meilleur portrait de la demande des patients actuels. En connaissant la fin des épisodes de soins pour chaque patient, le gestionnaire pourra déterminer d'avance l'allocation d'infirmières nécessaires pour combler la demande pour les soins de plaie dans le mois. De plus, puisque l'horaire est sur 31 jours, une erreur d'une vingtaine de jours comme c'est le cas, dans la majorité de nos modèles, est acceptable. Or, si le gestionnaire construisait son horaire à chaque semaine, une erreur d'une vingtaine de jours serait trop grande. Une autre retombée intéressante est lorsqu'un nouveau patient se présente à l'agence, le gestionnaire pourrait déterminer, en vérifiant la fin des épisodes de soins, si l'agence est capable de répondre à la demande de ce nouveau patient ou si toutes les infirmières ont déjà une charge à pleine capacité pour le mois. Dans ce cas, l'agence peut décider de refuser le patient ou encore d'engager une infirmière supplémentaire. Inversement, si un patient arrive et que le gestionnaire sait que les infirmières peuvent prendre en charge un nouveau patient, il pourrait l'accepter sans problème.

5.2 Analyse et discussion

La prochaine section est une analyse et une discussion sur le projet ainsi que sur les résultats obtenus. Plusieurs facteurs peuvent expliquer nos résultats et c'est ce que nous vous démontrerons dans cette partie du mémoire. De plus, des recommandations sont énoncées à la fin

de cette section.

5.2.1 Facteurs manquants

Présentement, la fréquence des visites ainsi que le nombre de visites réalisées auprès du patient sont déterminés selon le jugement clinique de l'infirmière. Tel qu'expliqué dans la méthodologie, s'il y a présence d'un proche, capable d'aider aux soins, il se peut que l'infirmière doive se présenter moins fréquemment au chevet de cette personne. Un proche n'est peut-être pas en mesure de détecter un signe précoce d'infection ou d'une problématique quelconque qui retarderait la guérison. Or, une infection peut rallonger significativement le temps de guérison d'une plaie.

Dans notre modèle, nous n'avons pas utilisé la variable qui indiquait le nombre de visites effectué au patient, bien qu'il puisque s'agir d'un facteur. Cette variable est inconnue à l'admission du patient. Ainsi, pour un nouveau patient, il est impossible de dire d'avance son nombre de visites, mais il peut s'agir, dans certains cas, d'un facteur très influençant.

De plus, de nombreux autres facteurs agissant sur le temps de guérison d'une plaie ne sont pas présents dans la base de données. En effet, selon la littérature, de nombreuses variables font varier le temps de guérison d'une plaie. Voici une liste des facteurs importants qu'il aurait été pertinent de posséder dans nos données :

1. Alimentation et hydratation
2. Anxiété
3. Poids
4. Tabagisme
5. Consommation d'alcool
6. Diagnostics médicaux
7. Présence de corps étranger dans la plaie
8. Soutien social
9. Croyances culturelles
10. Médication actuelle
11. Analyses sanguines

Certains de ces facteurs ont pu être trouvés dans les plans de soins, cependant il s'agit d'une minorité de patients qui présentent un plan de soins. De plus, pour les plans de soins, l'infirmière rédige ce qu'elle souhaite, et donc la grande majorité du temps, aucune information n'est donnée sur les facteurs énumérés précédemment. Il est donc fortement possible qu'un

patient possède certaines caractéristiques que nous avons cherchées dans le plan de soins, mais que l'infirmière ne l'ait simplement pas noté.

5.2.2 Valeurs manquantes

Un autre facteur qui a certainement influencé nos résultats concerne le nombre important de valeurs manquantes. Pour l'âge, par exemple, chez 1666 plaies il n'y avait aucune valeur pour cette variable. Comme pour cette variable, le nombre était très grand, nous avions décidé de remplacer les valeurs manquantes par la moyenne, soit 66 ans. Ainsi, nous n'avions pas à rejeter toutes les observations avec cette valeur manquante. Cependant, il est certain qu'une telle modification influence nos résultats. De plus, comme nous n'étions pas en mesure de dire si cette variable était corrélée avec la variable cible, nous ne voulions pas la rejeter complètement.

De nombreuses autres observations ont été supprimées, car de les remplacer par la moyenne était trop incertain. De plus, comme nous avons décidé de séparer les observations par type de plaie, les échantillons pour trouver les meilleurs modèles n'avaient pas un grand nombre d'observations. Pour avoir un aperçu, dans le fichier test pour les modèles contenant l'ensemble des observations, il y avait 993 données et donc le fichier d'apprentissage contenait 2317 données. Ce nombre est déjà petit pour un projet de ce type. Donc pour les autres plaies ayant un échantillon beaucoup plus petit, le nombre d'observations est encore plus minime, ce qui a grandement influencé les résultats. Nous avions donc un échantillon de données trop petit pour obtenir un résultat satisfaisant.

De plus, dans le cadre de notre projet, nous avons utilisé les observations de la première visite d'une infirmière chez un patient ayant une plaie, car nous désirions prédire à la première rencontre. Cependant, nous avions en notre possession les valeurs de chaque caractéristique lorsque les visites de l'infirmière avaient été réalisées. En analysant ces valeurs, on se rend compte que de nombreuses fois l'infirmière n'a pas inscrit des valeurs. En effet, sur toutes les visites faites à domicile, seulement 12% du temps une valeur est inscrite pour la variable superficie, par exemple.

Toujours concernant les valeurs manquantes, il semble qu'à plusieurs reprises les diagnostics des patients ne furent pas entrés dans le logiciel. En effet, au Canada en 2014, près de 50% des hommes et près de 55% des femmes âgées de 75 ans et plus souffraient d'hypertension artérielle [51]. Dans nos résultats, chez les patients âgés de 75 ans et plus, moins de 20% présentaient un diagnostic d'hypertension artérielle. Une différence de plus de 30%. Même son de cloche pour d'autres caractéristiques tel que l'obésité. Sur l'ensemble des patients présents dans notre base de données finale, seulement 40 personnes présentaient de l'obésité

soit moins de 7%, lorsqu'encore une fois les statistiques pour 2014 indiquent que plus de 20% de la population aurait un problème d'obésité [50]. Nous pouvons donc douter que l'ensemble des diagnostics des patients fût entré correctement dans la base de données.

De plus, pour l'ulcère diabétique du pied, seulement 44% des patients ayant ce type de plaie, possèdent un diagnostic de diabète inscrit dans la base de données. Ce qui est impossible puisque la définition d'un ulcère diabétique du pied signifie un ulcère chez un patient diabétique [30]. Ainsi, 100% de nos patients avec ce type de plaie auraient dû avoir un diagnostic de diabète. Il s'agit donc d'une autre preuve qu'il y a plusieurs erreurs dans la base de données.

5.2.3 Détection valeurs aberrantes

Un autre facteur qui a influencé nos résultats concerne la grande difficulté de détecter les valeurs aberrantes, c'est-à-dire des erreurs qui auraient été commises dans la prise des observations. En effet, parfois les mesures pour les dimensions de la plaie n'étaient peut-être pas exactes. De plus, selon la littérature, les dimensions d'une plaie sont des facteurs très importants. Or, il semble avoir plusieurs erreurs dans la base de données à ce sujet, certaines observations ont pu être rejetées grâce à notre analyse, cependant de nombreuses autres n'ont pu être décelées.

5.2.4 Bruit dans les données

Il semble y avoir du bruit dans nos données. En *data mining*, lorsque nous disons que nos données sont bruitées, cela signifie qu'il y a des erreurs dans les données souvent causées par de mauvaises prises d'observations.

En effet, nous avons analysé spécifiquement certains patients pour voir l'évolution de leur plaie visite après visite. Nous avons été étonnés de constater qu'à certains moments une valeur beaucoup plus grande était inscrite dans un court moment.

Par exemple, pour une certaine plaie, nous remarquons que la superficie de la plaie est relativement petite au départ puis augmente. L'évolution de cette plaie nous laisse perplexes. De nombreuses plaies que nous avons analysées brièvement montrent des évolutions pour la superficie de la plaie, ce qui nous amène à nous interroger sur la validité de celles-ci. Donc il y a de fortes chances que certaines plaies ne présentent pas des mesures précises, ce qui influence nécessairement nos résultats.

La figure 5.1 démontre l'exemple d'une certaine plaie ayant une évolution ambiguë selon une infirmière en soins de plaies.

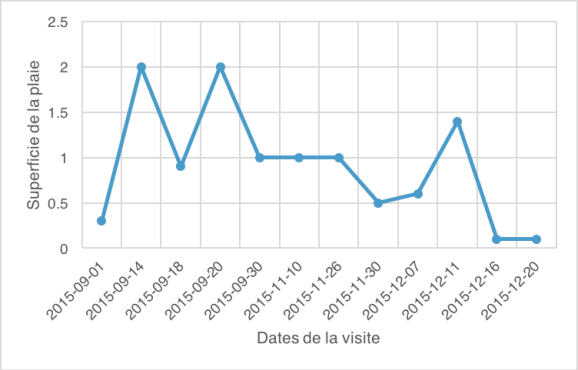

Figure 5.1 Exemple superficie de la plaie dans le temps

Concernant les modèles que nous utilisons, il y a beaucoup de paramètres, ce qui fait qu'il est presque impossible d'arriver à une réponse optimale, c'est-à-dire de trouver le taux d'erreurs le plus bas ou encore le taux de bonne classification le plus élevé.

De nombreux aspects nous poussent à croire qu'il y a du bruit dans les données. En effet, lorsque nous modifions le *random state*, c'est-à-dire lorsque nous prenons un échantillon différent pour nos fichiers, le taux d'erreurs change complètement. Parfois, il est bien meilleur et à d'autres moments il est pire. Il est cependant impossible de savoir quel modèle nous donnerait les meilleurs résultats avec un tout nouveau échantillon.

Habituellement, lorsque nous modifions le *random state* pour un fichier ayant des données propres, les taux d'erreurs ne devraient pas varier significativement.

5.2.5 Paramètres utilisés dans les modèles

De plus, pour les arbres de décision, il faut choisir une profondeur maximale pour l'arbre. C'est-à-dire le nombre de niveaux à partir du nœud racine. Nous avons testé de nombreuses valeurs différentes pour trouver les meilleurs arbres, cependant il était impossible de toutes les tester, et ce par manque de temps. Considérant cette contrainte, nous avons décidé de changer seulement ce paramètre pour les arbres de régression. Cependant, de nombreux autres paramètres étaient disponibles. Par exemple, nous aurions pu modifier plus de sept autres paramètres, tels que le nombre minimal d'observations nécessaire pour faire une séparation.

Même son de cloche pour les forêts aléatoires où plusieurs paramètres peuvent être remaniés, mais où nous avons seulement modifié certains paramètres.

5.2.6 Techniques utilisées pour détecter la présence de sur-apprentissage

Dans le cadre de ce projet, nous avons utilisé deux fichiers distincts, un pour l'apprentissage du modèle et l'autre pour tester le modèle afin de détecter la présence de sur-apprentissage. L'utilisation de deux échantillons permet de détecter la présence de sur-apprentissage. Cependant, en faisant ainsi, nous diminuons significativement notre nombre d'observations, dans notre cas de 30%, puisque nous avons utilisé un fichier test contenant ce pourcentage de toutes nos observations. Notre fichier d'apprentissage est donc relativement petit lorsque nous entraînons notre modèle.

Pour les modèles de type régression, nous avons accepté une différence maximale de six jours entre le fichier d'apprentissage et celui de test. Pour les modèles de type classification, une différence inférieure à 6% a été jugée acceptable pour un échantillon de la grosseur que nous utilisons. Le risque de sur-apprentissage était négligeable avec une telle différence. Or, il se peut que le pourcentage ne soit pas l'optimal.

5.2.7 Séparation des classes pour les modèles de classification

Toujours pour les modèles de classification, un autre facteur important qui a joué sur nos résultats est la séparation de nos classes. En effet, les intervalles ont été choisis en observant la distribution des valeurs pour la variable cible soit le nombre de jours nécessaires pour la guérison d'une plaie. Un résumé de la démarche pour le choix des intervalles est énoncé dans la partie méthodologie. La méthode utilisée pour le choix des intervalles était exploratoire et donc nous aurions pu continuer pour trouver d'autres intervalles pertinents.

La figure 5.2 représente la distribution générale de notre base de données initiale.

Une autre technique que nous aurions pu utiliser est l'analyse de ce tableau dans le choix de nos intervalles. Par exemple, si nous désirons créer seulement deux intervalles. La séparation pourrait être faite à 45 jours, car il s'agit de la médiane de notre distribution. Ainsi la moitié de toutes nos observations se retrouve avant ce nombre de jours et l'autre après. Si nous désirions déterminer plus d'intervalles, nous aurions pu pousser davantage l'analyse de cette figure.

5.2.8 Corrélation entre les variables explicatives et la variable cible

Dans la partie méthodologie, nous avions vérifié les coefficients de corrélation de toutes les variables explicatives disponibles. Il avait été décidé de garder l'ensemble des variables, sauf la superficie, car celle-ci était le résultat d'une équation entre deux autres variables explicatives.

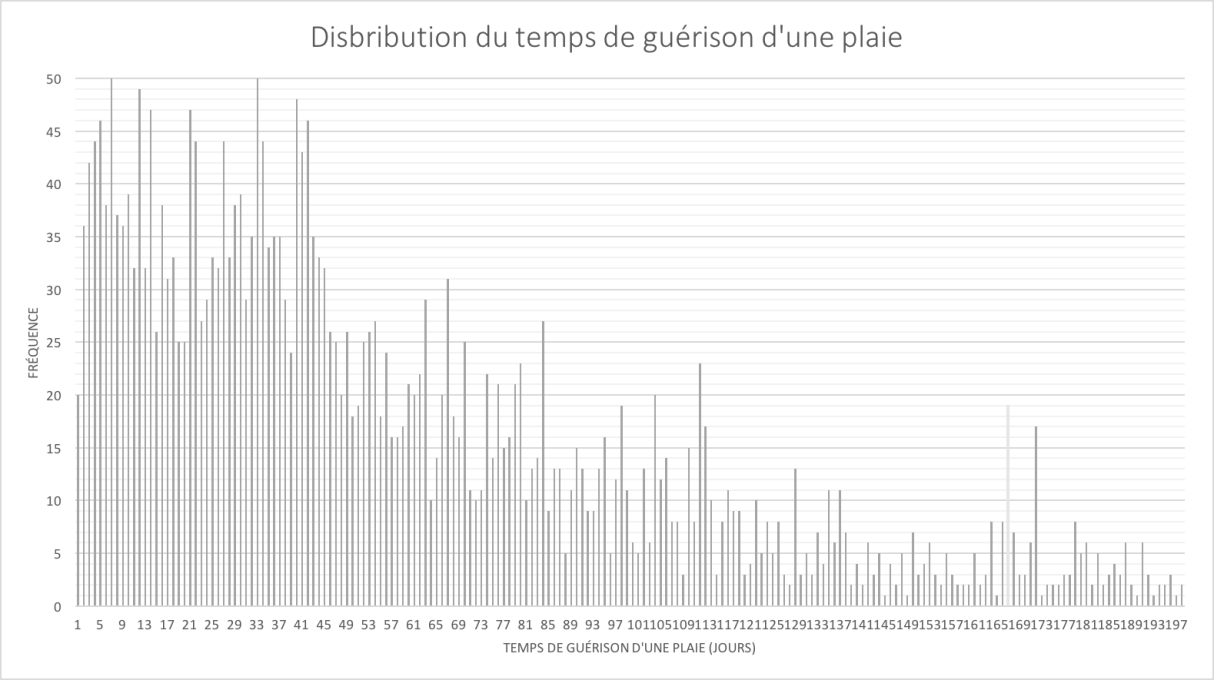

Figure 5.2 Distribution des temps de guérison d'une plaie

Nous remarquions cependant que plusieurs aberrations étaient présentes. En effet, prenons l'exemple de la variable explicative obésité, 8 fois sur 10, le coefficient de corrélation était négatif, ce qui signifie que présenter cette caractéristique soit être obèse diminuait le temps de guérison d'une plaie. De plus, la moyenne de ce coefficient est de -0.46, ce qui représente notre valeur la plus près de -1 ou de 1 pour un coefficient positif. Or, dans la littérature, il est indiqué que l'obésité est un des facteurs qui retarde le plus le temps de guérison d'une plaie.

Même interrogation pour la variable diabète où le coefficient de corrélation est aussi négatif. On sait cependant qu'en raison de la physiopathologie du diabète, le temps de guérison d'une plaie s'avère plus long. Toutefois, ce n'est pas ce que nous indique le coefficient moyen et où encore 8 fois sur 10 dans les essais, ce coefficient est négatif.

Plusieurs raisons peuvent expliquer un tel phénomène. Prenons l'obésité, par exemple, c'est en raison des autres facteurs et du nombre beaucoup plus petit des observations contenant un diagnostic d'obésité qu'on se retrouve avec de tels résultats. Dans notre base de données finale, seulement 40 personnes présentent cette caractéristique. De plus, lorsque nous comparons les résultats du temps de guérison d'une plaie chez un patient diabétique et un non-diabétique, nous ne prenons pas en compte les autres variables explicatives. Par exemple, il se peut que toutes les personnes présentant de l'obésité avaient une plaie plus petite que la moyenne des

non obèses et donc ceci peut impliquer un temps de guérison plus rapide.

5.2.9 Variabilité des données

Ce qui nous amène à parler de la variabilité de nos données. Dans notre base de données, il y a beaucoup de variabilité, c'est-à-dire qu'il y a de nombreuses combinaisons possibles entre nos variables explicatives. Pour la majorité des cas, il n'y a jamais deux combinaisons pareilles. Il s'agit d'une des raisons pour laquelle les modèles ont donc beaucoup de difficulté à prédire correctement.

5.2.10 Recommandations

Tel qu'énoncé dans les paragraphes précédents, plusieurs failles ont été découvertes dans la base de données. Certaines recommandations peuvent donc être faites en lien avec les limites présentées. En effet, nous croyons que de meilleurs résultats pourraient être obtenus si les recommandations que nous proposons étaient mises en place.

Il serait très important de revoir la prise de données, en effet nous croyons que de meilleurs résultats auraient pu être obtenus s'il n'y avait pas autant de valeurs manquantes. Nous recommandons donc à la compagnie *AlayaCare* la création de paramètres additionnels dans sa base de données. En effet, ceci aiderait l'infirmière à se souvenir de tous les points importants à inscrire dans ses notes. De plus, il est beaucoup plus facile et rapide, pour l'analyste ainsi que pour l'infirmière, de cocher des informations dans le logiciel que de tout écrire dans les plans de soins. Le logiciel d'*AlayaCare* pourrait aussi contenir une alarme lorsqu'il y a présence d'une valeur manquante au moment où l'infirmière quitte l'application. Ainsi, nous pouvons supposer qu'une baisse significative des valeurs manquantes aurait lieu. De plus, selon nous, la qualité des observations serait augmentée si une telle modification était implantée. Une rencontre pourrait aussi être faite avec les infirmières responsables dans les agences d'infirmières où les données sont récoltées, ainsi les infirmières pourraient comprendre l'importance d'entrer correctement l'ensemble des observations qu'elles font avec les patients.

De plus, il aurait été intéressant de tenter plusieurs autres groupes d'intervalles pour les modèles de classification. Nous aurions pu tenter la technique énoncée dans la section 5.2.7. De meilleurs résultats auraient peut-être pu être obtenus de cette manière.

Une autre amélioration future concerne le choix des variables. D'autres techniques telles que la méthode pas à pas, auraient pu être expérimentée pour déterminer quelles variables explicatives possèdent une corrélation avec la variable cible. En effet, en ayant moins de variables explicatives, il risque d'avoir moins de variabilité dans la base de données puisqu'il

y a moins de combinaisons possibles.

Dans le même ordre d'idées, plusieurs autres modifications des paramètres pour les modèles utilisés auraient pu être tentées. En effet, nous pourrions utiliser de nombreuses autres combinaisons de paramètres et même, nous aurions pu travailler sur d'autres paramètres présents dans les librairies utilisées.

CHAPITRE 6 CONCLUSION

Au cours de ce mémoire, nous avons présenté plusieurs modèles de prédiction, certains de type régressif et d'autres de classification. Nous avons été en mesure de prédire le temps de guérison d'une plaie en fonction de nombreuses variables explicatives, et ce pour trois types de plaies spécifiques, mais aussi pour l'ensemble des plaies. Nous avons donc répondu à notre objectif qui était de prédire le temps de guérison d'une plaie.

Dans le but d'obtenir les meilleurs résultats possible, de nombreux modèles ont été testés. Voici la liste des modèles que nous avons utilisés :

1. Régression linéaire
2. Régression logistique
3. Arbre de décision de type régression
4. Arbre de décision de type classification
5. Forêt aléatoire de type régression
6. Forêt aléatoire de type classification

De plus, pour les arbres et les forêts aléatoires, nous avons testé plusieurs combinaisons de paramètres, mais en s'assurant qu'il n'y ait pas de sur-apprentissage. Pour ce faire, nous avons travaillé avec deux techniques, soit l'utilisation de deux fichiers distincts ainsi que la validation croisée. Nous avons noté de meilleurs résultats avec la première technique, soit l'utilisation de fichiers distincts. En utilisant ces techniques, nous pouvions ainsi détecter la présence de sur-apprentissage. Dans le cas échéant, nous avons testé de nouveau notre modèle, mais en modifiant certains paramètres.

Il faut se rappeler qu'avant d'appliquer les modèles, il est important de faire un inventaire des données existantes ainsi qu'un travail minutieux de préparation des données. Dans cette phase de préparation des données, de nombreuses étapes sont nécessaires telles que la modification des valeurs manquantes.

La moyenne observée du temps de guérison d'une plaie est de 64,45 jours, et ce pour l'ensemble des plaies. Pour les résultats, en utilisant les algorithmes de régression, nous trouvons un nombre de jours d'erreurs moyens de 27,56 jours. Pour les résultats de classification, nous avons un taux de bonne classification moyen de 74,32% pour le groupe 1, 64,92% pour le groupe 2 et de 55,99% pour le groupe 3. Les intervalles des groupes sont énumérés à la section 4.3.3. Pour trouver ces résultats, nous avons modifié à plusieurs reprises les paramètres qui se trouvaient dans les modèles utilisés. Nous tentions de découvrir les meilleurs résultats,

mais tout en éliminant le sur-apprentissage dans nos modèles. Pour déterminer s'il y avait du sur-apprentissage dans nos modèles, nous avons utilisé deux fichiers distincts, car c'est la technique où l'on trouvait les meilleurs résultats.

Dans la discussion, plusieurs facteurs qui influencent négativement nos données ont été identifiés, tels que le bruit dans les données et certains facteurs absents influençant le temps de guérison d'une plaie. Nous sommes arrivés à la conclusion que la base de données utilisée ne permet pas de trouver des résultats avec de petites erreurs.

Nous avons suggéré quelques recommandations à la compagnie *AlayaCare*, celles-ci se retrouvent à la section 5.2.10. Une d'entre elles, concerne l'importance de revoir la prise d'informations et d'ajouter des alarmes pour insister les infirmières à bien remplir toutes les cases demandées. Présentement, il serait possible pour l'agence d'infirmières d'utiliser nos résultats pour s'aider dans la planification des horaires des infirmières, si leur horaire est construit sur un horizon de temps de plus de 31 jours, puisque nous avons en moyenne une vingtaine de jours d'erreurs. Nous croyons qu'avec une base de données plus complète et avec les recommandations que nous effectuons, un nombre de jours d'erreurs plus petit pourrait être obtenu. Nous croyons donc que ce projet ouvre la porte à d'autres travaux permettant l'optimisation des soins de plaies à domicile.

RÉFÉRENCES

[1] Altman, D. G. et J. M. Bland. 1994, «Statistics notes-diagnostic-tests-1-sensitivity and specificity. 3.», .

[2] Baraldi, R. «Coup d'œil sur les soins et services à domicile reçus par les aînés au québec en 2013-2014», URL `http://www.stat.gouv.qc.ca/statistiques/sante/bulletins/zoom-sante-201605.pdf`.

[3] Baranoski, S. et E. A. Ayello. 2008, *Wound care essentials : Practice principles*, Lippincott Williams & Wilkins.

[4] Barnston, A. G. 1992, «Correspondence among the correlation, rmse, and heidke forecast verification measures ; refinement of the heidke score», *Weather and Forecasting*, vol. 7, n° 4, p. 699–709.

[5] Beitz, J. M. 2012, «Predictors of success on wound ostomy continence nursing certification board examinations : A regression study of academic factors», *Journal of Wound Ostomy & Continence Nursing*, vol. 39, n° 4, p. 377–381.

[6] Bellavance, F. 2017, «Préparation de données pour le data mining», Cours universitaire HEC Montréal.

[7] Berry, M. J. et G. Linoff. 1997, *Data mining techniques : for marketing, sales, and customer support*, John Wiley & Sons, Inc.

[8] Campbell, M. J. 2001, «Multiple linear regression», *Statistics at Square Two : Understanding Modern Statistical Applications in Medicine, Second Edition*, p. 10–31.

[9] Canadian Healthcare Association. 2009, *Home care in Canada : From the margins to the mainstream*, Canadian Healthcare Association.

[10] Chapman, P., J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer et R. Wirth. 2000, «Crisp-dm 1.0 step-by-step data mining guide», .

[11] Clavet, N.-J., J.-Y. Duclos, B. Fortin, P.-C. Michaud et S. Marchand. 2013, *Les dépenses en santé du gouvernement du Québec, 2013-2030 : projections et déterminants*, Centre interuniversitaire de recherche en analyse des organisations.

[12] Clinic, W. C. 2013, «High blood pressure and wound care», .

[13] Cousineau, D. «Cours 12 : Corrélation et régression», .

[14] Dangare, C. S. et S. S. Apte. 2012, «Improved study of heart disease prediction system using data mining classification techniques», *International Journal of Computer Applications*, vol. 47, n° 10, p. 44–48.

[15] Danjuma, K. et A. O. Osofisan. 2015, «Evaluation of predictive data mining algorithms in erythemato-squamous disease diagnosis», *arXiv preprint arXiv :1501.00607*.

[16] Delen, D., G. Walker et A. Kadam. 2005, «Predicting breast cancer survivability : a comparison of three data mining methods», *Artificial intelligence in medicine*, vol. 34, nº 2, p. 113–127.

[17] Dietterich, T. 1995, «Overfitting and undercomputing in machine learning», *ACM computing surveys (CSUR)*, vol. 27, nº 3, p. 326–327.

[18] Foucart, T. 2006, «Colinéarité et régression linéaire», *Mathématiques et sciences humaines*, , nº 173, p. 5–26.

[19] Frescos, N., R. Nay, D. Fetherstonhaugh et S. Gibson. 2011, «Assessment and management of pain in chronic wounds : a national survey of australian health care practitioners caring for people with chronic wounds», *Journal of Foot and Ankle Research*, vol. 4, nº 1, p. O16.

[20] Gallant, C. 2008, *Évaluation des connaissances et des pratiques des infirmières, en centre hospitalier universitaire, en matière de prévention et de traitement des plaies de pression*, thèse de doctorat, Université Laval.

[21] Gauthier, J., N. Thompson, L. Bouffard, P. Plourde, C. Roy, M. Lajoie et S. Sadler. , «La santé dans tous ses états : les déterminants sociaux de la santé», URL `https://www.csssbcstl.qc.ca/fileadmin/csss_bcsl/Menu_du_haut/Publications/Trousse_pedagogique/Trousse_pedagogique_finale.pdf`.

[22] Gauthier, J., N. Thompson, L. Bouffard, P. Plourde, C. Roy, M. Lajoie et S. Sadler. , «Une action concertée pour optimiser le traitement des plaies chroniques et complexes», URL `http://www.oeq.org/DATA/NORME/24~v~cadre-de-collaboration-interprofessionnelle.pdf`.

[23] Gould, L., P. Abadir, H. Brem, M. Carter, T. Conner-Kerr, J. Davidson, L. DiPietro, V. Falanga, C. Fife et S. Gardner. 2015, «Chronic wound repair and healing in older adults : current status and future research», *Wound Repair and Regeneration*, vol. 23, nº 1, p. 1–13.

[24] Grégoire, G. 2014, «Multiple linear regression», *European Astronomical Society Publications Series*, vol. 66, p. 45–72.

[25] Guo, S. et L. A. DiPietro. 2010, «Factors affecting wound healing», *Journal of dental research*, vol. 89, nº 3, p. 219–229.

[26] Heikes, K. E., D. M. Eddy, B. Arondekar et L. Schlessinger. 2008, «Diabetes risk calculator», *Diabetes care*, vol. 31, nº 5, p. 1040–1045.

[27] Hess, C. T. 2011, «Checklist for factors affecting wound healing», *Advances in skin & wound care*, vol. 24, nº 4, p. 192.

[28] Hoover, M. et M. Rotermann. 2012, *Le recours aux soins à domicile par les personnes âgées et les besoins insatisfaits, 2009*, Statistique Canada.

[29] Jaggi, Tarteaut, Donnat et Arbona. 2010, «L'impact des odeurs dans les soins de plaies», .

[30] Jeffcoate, W. J. et K. G. Harding. 2003, «Diabetic foot ulcers», *The lancet*, vol. 361, nº 9368, p. 1545–1551.

[31] Lakshmi, M. S., D. Haritha et V. SRKIT. 2016, «Heart disease diagnosis using predictive data mining», *International Journal of Computer Science and Information Security*.

[32] Lazarus, G. S., D. M. Cooper, D. R. Knighton, D. J. Margolis, R. E. Pecoraro, G. Rodeheaver et M. C. Robson. 1994, «Definitions and guidelines for assessment of wounds and evaluation of healing», *Archives of dermatology*, vol. 130, nº 4, p. 489–493.

[33] LeBlanc, K., S. Baranoski, D. Christensen, D. Langemo, K. Edwards, S. Holloway, M. Gloeckner, A. Williams, K. Campbell, T. Alam et collab.. 2016, «The art of dressing selection : a consensus statement on skin tears and best practice», *Advances in skin & wound care*, vol. 29, nº 1, p. 32–46.

[34] Lee, S.-M., J.-O. Kang et Y.-M. Suh. 2004, «Comparison of hospital charge prediction models for colorectal cancer patients : neural network vs. decision tree models», *Journal of Korean medical science*, vol. 19, nº 5, p. 677–681.

[35] Legrand, P. et D. Bories. «Le choix des variables explicatives dans les modèles de régression logistique», .

[36] Lemberger, P., M. Batty, M. Morel et J.-L. Raffaëlli. 2015, *Big Data et machine learning : Manuel du data scientist*, Dunod.

[37] Lessard, J., A. Blancquaert et M. Bernard. 2017, «Budget du québec 2017-2018», URL `http://www.aviseo.ca/assets/budget-qc-2017-2018-vf.pdf`.

[38] Mackavey, C. 2016, «Advanced practice nurse transitional care model promotes healing in wound care», *Care Management Journals*, vol. 17, nº 3, p. 140–149.

[39] Marchildon, G. P. 2011, *Health care cost drivers : the facts*, Canadian Institute for Health Information.

[40] Maroco, J., D. Silva, A. Rodrigues, M. Guerreiro, I. Santana et A. de Mendonça. 2011, «Data mining methods in the prediction of dementia : A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression,

neural networks, support vector machines, classification trees and random forests», *BMC research notes*, vol. 4, n° 1, p. 299.

[41] Meng, X.-H., Y.-X. Huang, D.-P. Rao, Q. Zhang et Q. Liu. 2013, «Comparison of three data mining models for predicting diabetes or prediabetes by risk factors», *The Kaohsiung journal of medical sciences*, vol. 29, n° 2, p. 93–99.

[42] Nagelkerke, N. J. 1991, «A note on a general definition of the coefficient of determination», *Biometrika*, vol. 78, n° 3, p. 691–692.

[43] Palaniappan, S. et R. Awang. 2008, «Intelligent heart disease prediction system using data mining techniques», dans *Computer Systems and Applications, 2008. AICCSA 2008. IEEE/ACS International Conference on*, IEEE, p. 108–115.

[44] Peng, C.-Y. J., K. L. Lee et G. M. Ingersoll. 2002, «An introduction to logistic regression analysis and reporting», *The journal of educational research*, vol. 96, n° 1, p. 3–14.

[45] Saadoun, M. «I–du data warehouse au data mart», .

[46] Sarvestani, A. S., A. Safavi, N. Parandeh et M. Salehi. 2010, «Predicting breast cancer survivability using data mining techniques», dans *Software technology and Engineering (ICSTE), 2010 2nd international Conference on*, vol. 2, IEEE, p. V2–227.

[47] Seaman, S. 2005, «The role of the nurse specialist in the care of patients with diabetic foot ulcers», .

[48] Sheehan, P., P. Jones, A. Caselli, J. M. Giurini et A. Veves. 2003, «Percent change in wound area of diabetic foot ulcers over a 4-week period is a robust predictor of complete healing in a 12-week prospective trial», *Diabetes care*, vol. 26, n° 6, p. 1879–1882.

[49] Snyder, R. J., C. Fife et Z. Moore. 2016, «Components and quality measures of dime (devitalized tissue, infection/inflammation, moisture balance, and edge preparation) in wound care», *Advances in skin & wound care*, vol. 29, n° 5, p. 205.

[50] Statistiques Canada. , «Diabète 2014», URL `http://www.statcan.gc.ca/pub/82-625-x/2015001/article/14180-fra.htm`.

[51] Statistiques Canada. , «Hypertension artérielle 2014», URL `http://www.statcan.gc.ca/pub/82-625-x/2015001/article/14184-fra.htm`.

[52] Stel, V. S., S. M. Pluijm, D. J. Deeg, J. H. Smit, L. M. Bouter et P. Lips. 2003, «A classification tree for predicting recurrent falling in community-dwelling older persons», *Journal of the American Geriatrics Society*, vol. 51, n° 10, p. 1356–1364.

[53] Swift, M. E., A. L. Burns, K. L. Gray et L. A. DiPietro. 2001, «Age-related alterations in the inflammatory response to dermal injury», *Journal of Investigative Dermatology*, vol. 117, n° 5, p. 1027–1035.

[54] Takahashi, P. Y., L. J. Kiemele et J. P. Jones. 2004, «Wound care for elderly patients : advances and clinical applications for practicing physicians», dans *Mayo Clinic Proceedings*, vol. 79, Elsevier, p. 260–267.

[55] Taneja, A. 2013, «Heart disease prediction system using data mining techniques», *Oriental Journal of Computer science and technology*, vol. 6, n° 4, p. 457–466.

[56] Thuraisingham, B. M. et M. G. Ceruti. 2000, «Understanding data mining and applying it to command, control, communications and intelligence environments», dans *Computer Software and Applications Conference, 2000. COMPSAC 2000. The 24th Annual International*, IEEE, p. 171–175.

[57] Tortora, G. et S. Grabowski. «Principles of anatomy and physiology, edited by bonnie roesh», .

[58] Vézina, J. et C. Saint Pierre. 2006, «Cahier 7 le soin des plaies : Principes de bases», .

[59] Wang, S. B., K. M. Hu, K. J. Seamon, V. Mani, Y. Chen et K. Gronert. 2012, «Estrogen negatively regulates epithelial wound healing and protective lipid mediator circuits in the cornea», *The FASEB Journal*, vol. 26, n° 4, p. 1506–1516.

[60] Wound Care Center. «Different types of wounds», URL `http://www.woundcarecenters.org/article/wound-basics/different-types-of-wounds`.

ANNEXE A VALEURS POUR NORMALISATION

Les tableaux présents de l'annexe 1 représente les valeurs à utiliser pour normaliser les données des variables continues. Chaque tableau représente un ensemble de plaies. Ainsi, il faut utiliser les valeurs présentés avec le bon échantillon.

Tableau A.1 Variables à normaliser ; pour l'ensemble des plaies

Variable	**Moyenne**	**Écart type**
Âge	69.93	15.61
Longueur	3.37	1.88
Largeur	2.34	1.88
Profondeur	0.48	0.68
Superficie	9.04	9.49

Tableau A.2 Variables à normaliser ; pour les plaies opératoires

Variable	**Moyenne**	**Écart type**
Âge	63.52	14.67
Longueur	3.34	2.04
Largeur	2.05	1.41
Profondeur	0.96	0.95
Superficie	7.65	9.52

Tableau A.3 Variables à normaliser ; pour les plaies traumatiques

Variable	**Moyenne**	**Écart type**
Âge	72.39	15.64
Longueur	3.56	2.03
Largeur	2.34	1.45
Profondeur	0.22	0.30
Superficie	9.46	10.22

Tableau A.4 Variables à normaliser ; pour les plaies de pression

Variable	**Moyenne**	**Écart type**
Âge	74.40	16.21
Longueur	3.22	1.58
Largeur	2.49	1.44
Profondeur	0.58	0.76
Superficie	9.00	8.49